

科学技術用語の英日翻訳規則の抽出

5H-5

田上 文俊[†] 坂口 秀幸[‡] 竹田 正幸[†] 松尾 文碩[†][†]九州大学工学部 [‡]NEC 通信システム九州ソフトウェア部

1. まえがき

科学技術分野では、研究の進展に伴い次々と新たな専門用語がつくられる。したがって、科学技術文献の機械翻訳において、対訳辞書だけで対応するのは困難である。そこで、辞書に未登録の用語の英日自動翻訳法を開発するために、英日対訳辞書から翻訳規則の抽出を試みた。

2. 訳の抽出

まず、対訳辞書より、複数単語からなる語句の訳語と単一語句の訳語をもとに、新たな単語の訳語を抽出した。例えば、'aerial pollution'「大気汚染」と'pollution'「汚染」があれば'aerial'「大気」を抽出する。

今回、科学技術術語とその対訳のデータとして市販の機械可読である科学技術用語辞書(日外アソシエーツ社 EB科学技術用語大事典)を用いた。はじめに、辞書に登録されていた1単語からなる見出し語とその訳の対の集合を T_0 とする。例えば、(voltage, 電圧) $\in T_0$ である。単語に対する新たな訳を得るために以下のような処理を繰り返す。

i 回目の処理において、最初に $T_i = T_{i-1}$ とし、以後の作業は次のように行う。

- speed control system が和訳「速度制御装置」をもち、

(speed, 速度) $\in T_{i-1}$ (control, 制御) $\in T_{i-1}$ (system, 装置) $\notin T_{i-1}$

である場合、(system, 装置) を T_i に加える。すなわち、'system' に訳「装置」を追加する。

- working voltage が和訳「電圧定格」をもち、
(working, 定格) $\notin T_{i-1}$

(voltage, 電圧) $\in T_{i-1}$

である場合、(working, 定格) を T_i に加える。この場合は英単語の語順とその対訳の語順が異なるが、working に訳「定格」を追加する。

- arc resistance test が和訳「耐アーク性試験」をもち、

(arc, アーク) $\in T_{i-1}$ (test, 試験) $\in T_{i-1}$

である場合、(resistance, 耐*性) を T_i に加える。

この作業を、訳の増加がなくなるまで繰り返す。

訳を抽出する場合、単純にサブストリングマッチを行うだけでは不備が生じる場合がある。代表的なのは、カタカナ訳の読みの揺らぎによる場合で、例えば、'reactor' には「リアクター」「リアクタ」の二つの読みが考えられる。これより、

'arithmetic reactor' が和訳「演算リアクタ」をもち、

(arithmetic, 演算) $\notin T_{i-1}$ (reactor, リアクター) $\in T_{i-1}$

である場合、カタカナの変化に対応して「リアクター」と「リアクタ」を同じ'reactor'の訳と見なし、(arithmetic, 演算) を T_i に加えなければならない。

また次のような場合もある。'memory-mapped I/O' には「メモリマップ I/O」と「メモリマップド I/O」のふたつの訳が記述されている。ここで、'memory-mapped' 部分の訳には、原形'memory-map'の読みが使われている。このように英単語の語尾変化が日本人にとって発音しにくい場合など、日本人にとってより馴染みの深い、単語の原形などの読みが用いられる場合がある。

3. 抽出した訳の判定法

上の方法で生成した抽出訳の中には、'cellular'「自動車」という訳があった。これは、'cellular phone'の和訳「自動車電話」と'phone'の訳「電話」より抽出されたものである。ここで、'cellular phone'とは意識であり、'cellular'という英単語に対する「自動車」と

Extraction of English-Japanese Translation Rules for Technical Terms

Fumitoshi Tanoue[†], Hideyuki Sakaguchi[‡], Masayuki Takeda[†] and Fumihiro Matsuo[†]

[†]Kyushu University 36, Hakozaki, Fukuoka, 812 Japan

[‡]NEC Communication System Kyushu

いう訳はふさわしくないと考えられる。‘cellular’の訳語として、常用語辞書中には「細胞の」「細胞質の」と記述してある。

また、‘high’という英単語が「高」という抽出訳をもっていたが、これは、‘high performance’の和訳「高性能」と‘performance’の訳「性能」より抽出されたものである。この場合の「高」という抽出訳は、‘high’の和訳としてふさわしいと考えられる。‘high’の訳語としては、常用語辞書中に「高い」「高度な」などが記述されている。

このように、新たに抽出した訳と辞書の訳中に存在する訳語を比較し、1字でも一致した場合のみ、この抽出訳が使用可能だと判定する。

上の例のように、漢字の場合はほとんどその1文字ごとを形態素として英単語と意味的対応をとることができるので、この比較方法が成り立つ。

漢字以外の文字を含む抽出訳については、少々話が複雑になる。‘adhesive」「接着ガーゼ」のような抽出訳の場合、「接着」「ガーゼ」のように漢字とカタカナの間で日本語形態素が分離できる。「ガーゼ」の部分をも1形態素とみなし、「接着」の部分と同様に辞書による判定もできる。このように、文字の種類を形態素の境目とする場合が多いが、‘silicofluoride」「ケイフッ化水素酸」のような抽出訳も存在する。さらに、‘prone」「を起こしやすい」のような抽出訳の場合は、「を」の部分はひらがな1文字で助詞だということが推定できるが、漢字に続くひらがなをどのように区切るかが問題となる。‘prone’の訳は、常用語辞書中に「傾向があって、(～し) やすくて」とある。こういった抽出訳は辞書情報のみで判定するのが困難である。

4. 漢字のみからなる抽出訳の判定結果

漢字のみで構成される抽出訳について、上の方法を用いて判定した。初めに、抽出訳の判定を、訳の抽出を行った科学技術用語辞書で行ったが、この場合、‘after」「後」という抽出訳を使用不可と判定した。これは科学技術用語辞書に‘after」「…の後ろに」という自明な対訳が存在しないためである。このような一般的で自明な漢字対訳も使用可能と判定するために、常用語辞書(研究社 新英和中辞典 第5版)を用いることにした。

逆に、判定を常用語辞書のみで行った場合、‘active」「反応性」「放射性」という抽出訳を使用不可と判定した。このような専門用語も使用可能と判定するためには、両方の辞書を併用することが望ましいことがわかる。

漢字抽出訳 22130 について判定を行った結果、科学技術用語辞書では 10063、常用語辞書では 9154、また両方の辞書を用いた場合 13051 の抽出訳を使用可能と判定した。

二つの辞書の併用による判定の結果、使用可能とされた抽出訳の中には、次のようなものが存在した。例として、‘arithmetic」という単語に対し「演算」「演算機構」「桁演算」などの訳がある。「演算機構」「桁演算」という訳は、組み合わせる相手を限定することになり、使用は困難である。しかし、この二つの訳はどちらも「演算」という文字列を含むことから、‘arithmetic」という英単語が「演算」という日本語形態素と直接対応関係を持っていると考えられ、また「機構」「桁」という形態素は、‘arithmetic」に間接的に対応する冗長な形態素と見なすことができる。この冗長な形態素の使用法は、今後の課題とする。

この判定法を用いた場合、辞書中に記述のない英単語については全て使用不能と判定される。代表的なものに固有名詞がある。例えば、‘Ito」「伊東」という人名の訳も抽出されたが、この訳は辞書による判定法では選択できない。

5. むすび

抽出した訳の中から、それぞれの単語にふさわしい訳のみを、辞書を用いて選択した。

また、抽出訳中に漢字とその他の文字が混入している場合については、調査を進め、それぞれについてのより良い訳の選択法を決定していく。

抽出訳を判定した後、実際に対訳を合成する場合に、複数の訳語中から最適な訳の組合せを選択する方法をさらに調査する。

なお、本研究は、一部文部省科学研究費補助金(#07558162)の援助により行った。

参考文献

- 1) 梅山英昭：和文科学技術用語の形態素解析，九州大学工学部電気工学科卒業論文，1994
- 2) 坂口秀幸：科学技術用語の英日翻訳に関する研究，九州大学大学院工学研究科電気工学専攻修士論文，1995