不要語リストを用いたRFC英和辞書作成過程における課題

森　理、本田善久、（大同工業大学）、小川清（名古屋市工業研究所）

1はじめに

名古屋市工研では、情報技術に関する実用的な研究を行なっている。RFC(request for comment)は、インタネットの規約に関する情報源である。このRFCによって、インタネットの新しい提案、実験方法が示されている。インタネットの通信プロトコルは、完全性を保証しておらず、実験段階のものが多いだけでなく、それぞれのプロトコル同士の関係も明確でない場合もある。そのため、RFC を正確に理解するために、英文の個々の単語を分析し、RFC の文書が新たに出た場合、理解を容易にするための英和辞書の作成を検討した。

2　用語辞書の作成

第一段階としてすべてのRFCから単語を抽出し用語辞書を作成した。用語を分析するために用語の出現回数を集計し資料を作成した。また、実際のRFCを翻訳しその中での用語の引用と用語辞書での定義とを比較検討し、不要語辞書を作成することにより、新語検索を容易にすることとした。

具体的には、作業を開始した1999年4月1日現在のRFC中の単語一覧をTRコマンドとAWKKスクリプト、C プログラムを用いて作成し、単語の出現回数を集計した。

（1）単語を小文字に統一した。そのため大文字と小文字で意味が異なる場合も同一単語として集計した。これにより、言語的な操作を行わず、単純な処理だけでどのような単語リストが作成されるかを確認した。後に、固有名詞辞書を作成することにより、意味の分離を図った。

（2）（1）の処理だけでは、数字、記号の混ざった変数なども多く出現したため、数字、記号を削除した。(ピリオド、アポストロフィ、ハイフン、プラスなど)

（3）（2）の処理をしても、RFCがテキストにより図表を表示していることと、プログラム及びデータも表示しているため、3文字以上同じ続く単語が頻出した。このタイプの文字列が出現する回数が多いため、ノイズとして除去することとし、ひとまず同一の文字が3回以上連続する単語を削除した。(ZZZ など) ここまでは、単純処理の延長である。

（4）次に、副詞（ly)、動名詞（ing)、複数形（s、es)、過去形（d、ed）がすぐ近くにならぶため、これらの正規化を検討した。これらは、同一単語として集計した方が便利であるため、単純に加減演算で正規化できるものだけ処理した。非定形な活用のものはひとまず異なる単語として集計した。

作成した辞書は、表計算ソフトウェアで分析することとした。(1)のみの処理ではある表計算ソフトの容量を超えた。表計算ソフトで分析するために(2)、(3)の処理を行なった。

3　用語の出現回数の検討

当初、作成した用語辞書に含まれる単語は5万6千種類あった。すべてのRFCを翻訳していないので、変数の可能性のある用語のうち、短い単語の除外をすべては行っていない。これは、何が専門用語かの判断をする前に、推定で変数だと想定すると、専門用語、略語を除外する可能性があるからである。単語、熟語、文脈上の分析を行ないながら、不要語一覧を作り、削除するのが妥当であると考えた。

出現回数表では、the をはじめ、一般用語が多い。一般用語の出現回数では最多のものでは約73万回、専門用語でも約6万3千回を数えている。一方で、一度しか出現しない一度しか出現しない単語は1万6千種類あった。そこでて、英和辞書に搭載されている単語は一般用語とし、英和辞書データを一般用語辞書とした。これとマッチングしないものを専門用語辞書とした。

頻度	単語
732156	The
603575	A
293594	Of
264156	To
226055	I
193456	Be
170905	In
140911	For
113381	This
91459	That

4 RFC における用語の検討と複数辞書

用語辞書には、英和辞典などから一つづつ単語に日本語に付与した。翻訳の過程で実際の RFC での用語の使い方を検討した。すでに日本語に翻訳されている文献では、コンピュータ関係用語はカタカナ語であったり、アルファベットを用いて、一般用語と容易に識別できることが多い。コンピュータ用語以外では、動詞の場合、専門用語は漢字の熟語、一般用語は漢字 1 文字とひらがなによる場合もある。しかし、英語では、専門用語も一般用語も同じ単語を使う場合が多い。例えば、service は、特定のプロトコルが行う「サービス」である場合と、一般用語として助けの意味で使われていることもある。

頻度	専門用語
63741	Protocol
54676	Network
50029	Message
40663	Internet
37620	Information
36815	Object
32616	Standard
30448	Packet
29151	System
28782	Request

同一の単語でありながらまったく異なる定義がなされている場合がある。例えば、object のように、一般用語と専門用語として使われるほか、コンピュータの内部のものをさす場合と、コンピュータの外部のものをさす場合がある。その逆に異なる単語が同一の意味として使われているものもある。この場合は、出現数の多い単語で置き換えが可能であると考えられる。

出現回数	種類
1	16057
2	8433
3	4365
4	3645
5	2207
6	1986
7	1295
8	1196
9	1046
10	815

一般用語辞書、専門用語辞書、不要語辞書以外に、多くの副次的な辞書を作成した。電子メールアドレスのみを抽出した電子メール辞書は、文書名と電子メールアドレスにより、連絡先を一覧し、関連文書を電子メールアドレスから関連付けた。略号、固有名詞辞書は、大文字小文字の違いによる意味の違いと、略号とフルスペルの関係を明確にした。

まとめ

本研究では、RFC における用語辞書の作成とともに用語定義を行う上での資料の作成を行った。特に、不要語辞書は、手作業で行うが、一つの RFC が追加された場合には、人手で判断すべき単語数は多くないため、人手の判断で十分である。今後は、OS、プログラミング言語に固有な用語、通信プロトコルに固有な用語を分離し、RFC の理解を容易にできるようにすることを検討中である。

参考文献

[1] プロセス工程診断翻訳辞書 、小川清 h11 前期情報処理学会全国大会

[2] WRAPL による機械翻訳のためのテクスト編集と翻訳精度の向上、佐良木昌

[3] インターネットＲＦＣ辞典、笠野英松監修、マルチメディア通信研究会、アスキー出版局