対訳マニュアル文書からの対訳語の自動抽出

韓 暁峰*，松原 茂樹，吉川 正俊 (名古屋大学)

Automatic Extraction of Bilingual Words from Open-Source Software Manual

Xiaofeng Han, Shigeki Matsubara, Masatoshi Yoshikawa (Nagoya University)

1 はじめに

Linux や FreeBSD などのオープンソースソフトウェアのマニュアル文書は，多くの場合，英語で書かれており，日本語のサポートは必ずしも十分ではない．ソフトウェアマニュアルでは，文書特有の言い回しが多数現れるため，文書翻訳には，対訳事例を参照した処理が有効である．

本稿では，ソフトウェアマニュアルの自動翻訳に利用することを目的に，対訳マニュアル文書からの対訳語の抽出について述べる．対訳文書における単語の出現に関する統計情報を用いることにより，対訳辞書を使うことなく，単語対応を与えることができる．Red Hat Linux 9 の英日マニュアル文書を用いて対訳対応付け実験を行った．

2 対訳マニュアル文書から対訳語の自動抽出

本手法では，対訳マニュアル文書の文対応を推定し，対応する文の間での名詞の共起頻度に基づき，対訳語を推定する．

2.1 文対応の推定

Linux の対訳マニュアルの例を Fig.1 に示す．文書の特徴として以下があげられる．

1. 日本語と英語で文書構造が一致する．
2. 英語文と日本語文の文対応のうち，98.4%が 1 対 1，または，1 対 2 対応であり，訳文のスタイルは直訳である．
3. 関数など，マニュアル特有の名詞は，そのまま英語で表現される (Fig.1 における，“ closedir() ”など)．

そこで本研究では，文対応を推定するために，まず，マニュアル文書を文書要素に分割し (特徴 1)，次に，文書の先頭から順に，英語の 1 文を日本語の 1 文または 2 文に対応付ける (特徴 2)．そのとき，日本語文内のアルファベットの出現を手がかりとする (特徴 3)．

2.2 単語対応の推定

本研究では，対応する英語文と日本語文に共起する単語対は対訳対応にあるとして対訳語を抽出する [1]．すなわち，文対応に対して，英語文には Brill's tagger[2] を，日本語文には Chasen[3] を用いて形態素解析を実行し，英語文に出現するすべての名詞 w_e について，以下の Dice 係数を最大にする日本語名詞 w_j をその対訳語とする．

$$Dice(w_e, w_j) = \frac{2f_{ej}}{f_e + f_j}$$

NAME
closedir - close a directory
SYNOPSIS
#include <sys/types.h>
#include <dirent.h>
int closedir(DIR *dir);
DESCRIPTION
The closedir() function closes the directory stream associated with dir. The directory stream descriptor dir is not available after this call.
RETURN VALUE
The closedir() function returns 0 on success or -1 on failure.
ERRORS
EBADF Invalid directory stream descriptor dir. ……

名前
closedir - ディレクトリを閉じる
書式
#include <sys/types.h>
#include <dirent.h>
int closedir(DIR *dir);
説明
closedir()関数は dir に連結しているディレクトリストリームを閉じる。ディレクトリストリームディスクリプターdir は、この呼び出しの後では使用することができない。
返り値
closedir() 関数は成功時に 0 を返し失敗時に-1 を返す。
エラー
EBADF 無効なディレクトリストリームディスクリプターdir が呼ばれた。 ……

Fig. 1: Bilingual Text of Linux Manual(Closedir)

ここで，f_e は英語単語 w_e の出現頻度，f_j は日本語単語 w_j の出現頻度，f_{ej} は文対応における両単語の共起頻度である．

3 実験および評価

本手法の有効性を評価するため，Red Hat Linux 9 のマニュアル文書 man3 を用いて，対訳語推定実験を行った．実験では，328 対訳文書の英語 4613 文とその日本語訳文を用いた．文対応付けの正解率は 98.0%であった．対訳語推定の評価は，man3 の対訳マニュアルファイル xdr の 113 文対応に含まれる単語対応について，精度と再現率を計算することにより行った．

$$\text{精度} = \frac{\text{推定された対訳語のうち適合する対訳語の数}}{\text{推定された対訳語の数}}$$

$$\text{再現率} = \frac{\text{推定された対訳語のうち適合する対訳語の数}}{\text{全対訳語の数}}$$

実験の結果，精度 81.5%，再現率 77.9%であり，対訳マニュアル文書からの対訳語推定における本手法の有効性を確認した．

4 まとめ

本稿では，英日対訳ソフトウェアのマニュアル文書における対訳語推定手法について述べた．Linux の man3 を用いて，対訳語推定実験を行った．今後は，単語対応に基づくフレーズ対応付け手法について検討する予定である．

文献

(1) M.Ohara, et al: NLPKE, pp.150-157, 2003.
(2) E.Brill: Some Advance in Transformation-Based Part of Speech Tagging, AAAI, 1994.
(3) Y.Matsumoto, et al: Morphological Analysis System ChaSen version 2.2.1 Manual, 2000.