1N-9

オープンソースソフトウェアマニュアルを用いた対訳パターンの自動抽出

韓 暁峰† 松原 茂樹‡ 吉川 正俊†
† 名古屋大学大学院情報科学研究科 ‡ 名古屋大学情報連携基盤センター

1 はじめに

近年，Linux や FreeBSD などオープンソースソフトウェアの世界的な普及にともない，マニュアル文書の翻訳に対する需要が増加している．一般に，オープンソースソフトウェアマニュアルは大量に存在し，また，更新も頻繁に行われるため，人手による翻訳コストの軽減のために機械翻訳システムの活用が望まれる．マニュアル文書の翻訳は，多くの場合，直訳的で，言い回しも定型的であるため，翻訳処理においては対訳事例を利用したアプローチが有効である．

本稿では，ソフトウェアマニュアルの自動翻訳において利用するための，対訳マニュアル文書からの対訳パターンの抽出について述べる．Red hat Linux 9.0 の日本語版に付属する対訳マニュアル文書を使用し，対訳パターンの獲得実験を行った．

2 対訳マニュアル文書の特徴

オープンソースソフトウェアマニュアルの対訳文は概して直訳的であり，表現や構文等のばらつきは小さい．Linux の対訳マニュアルの例を図 1 に示す．対訳文書の特徴として以下があげられる．

1. 段落等の文書構造，及び，文の生起順序が英語と日本語で一致する．
2. 英語文と日本語文の文対応のほとんどが 1 対 1，または，1 対 2 対応である．
3. 関数名など，マニュアルに特有の固有名詞は，そのまま英語で表現される (図 1 における，“ closedir() ” など)．

3 対訳パターンの自動構築

対訳パターンを抽出するために，まず文対応及び単語対応付けを行う．両言語文に対して依存解析を実行し，依存関係に基づく対訳パターンを生成する．

3.1 文対応の推定

マニュアル文書を対訳コーパスとして利用するために，文の対応付けを行う必要がある．本手法では，前節で説明したマニュアル文書の特徴に基づき，対訳マニュアル文書の文対応を推定する．

まず，マニュアル文書を文書要素に分割し，文書の先頭から順に (特徴 1 より)，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 より)．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 より)．

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。 ……

図 1: Linux 対訳マニュアル文書の例

3.2 単語対応の推定

対訳語の獲得は，前処理として，まず，EDR 対訳辞書に含まれている対訳語を抽出し，単語対を獲得する．EDR 対訳辞書に含まれていない名詞の対訳語に対しては，対訳共起に関する統計情報を計算する．すなわち，文対応に対して，英語文には Brill's tagger[1] を，日本語文には Chasen[2] 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．

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

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．

3.3 対訳情報使用の対訳パターンの自動獲得

対訳パターンの獲得は，次の 3 つのステップからなる．

1. 依存関係の抽出
英語文，日本語文のそれぞれに対して，依存関係を生成する (図 2)．英語文では Rasp[3] を用い

Automatic Extraction of Translation Patterns from Manual for Open Source Software
† Xiaofeng Han (han@dl.itc.nagoya-u.ac.jp)
‡ Shigeki Matsubara (matubara@itc.nagoya-u.ac.jp)
‡ Masatoshi Yoshikawa (yosikawa@is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University(†)
Information Technology Center, Nagoya University(‡)

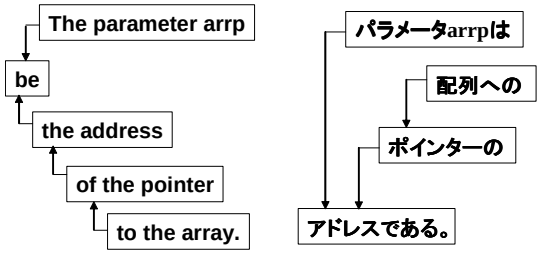

図 2: 英語と日本語文の依存構造

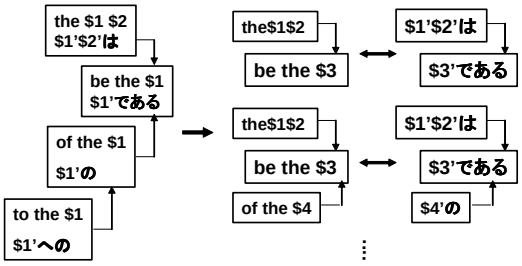

図 3: 対訳パターンの抽出

てチャンク間の依存関係を，また，日本語文では Cabocha[4] を用いて，文節間の係り受け関係を抽出する．なお，英語では，機能語は後続する内容語と，また，助動詞は主動詞とそれぞれまとめることにより，チャンクを作り上げる [5].

2. フレーズの対応付け
英語チャンクと日本語文節の対応付けのアライメントを行う (図 3 左)．まず，抽出された単語対と対訳文に含まれる英文字を用いて，英日のフレーズ間に対応関係を与える．未対応の部分に対し，依存構造に基づき，対応関係の推定を行う．この処理に必要なチャンクと文節の対応付けは，それぞれの構文構造の類似性により抽出する [6]．たとえば，未対応句とそれらの周辺の対応関係を考慮して，新規の対応付けと既存の対応を併合することにより，対応関係を推定する．
3. 対訳パターンの生成
抽出されたチャンクと文節との間の対応関係を対訳パターンの構築に利用する (図 3)．対訳パターンは英語チャンクと日本語文節間の対応関係の組合せにより，フレーズ単位から文単位まで対応付け，対訳句内の名詞語対をパラメータ化することによって対訳パターンを生成する．

4 対訳パターン抽出実験と評価

本手法の有効性を評価するために，対訳パターン抽出実験を行った．実験では，Red Hat Linux 9.0(日本語版) のマニュアル文書のうち，man3 の英日対訳ファイルを使用した．man3 は，ライブラリ関数について記したマニュアルであり，328 ファイルから構成されている．

本手法を Perl で実装した．対訳文推定で獲得した 4613 文対応に対して，対訳語推定で獲得した 1252 単語対応を用いて対訳パターンを抽出した．

評価は，xdr に関する対訳マニュアル文書に対して実施した．これは，英語 70 文，日本語 84 文からなり，man3 の中では rpc(英語 204 文) に続く規模のファイルである．文対応の正解率は 98.6%(69/70) であり，また，単語対応の精度と再現率はそれぞれ，98.4%(62/62)，95.4%(62/65) であった．

実験により，xdr ファイルから 389 対訳パターンを抽出することができた．事前に作成した正解データとの比較では，精度で 71.2%(277/389)，再現率で 50.6%(277/547) であり，本手法の実現可能性を確認した．

5 まとめ

本稿では，ソフトウェアマニュアルの対訳文書を用いた対訳パターンの構築手法について述べた．依存関係に基づいて対応付けを行うことにより，汎用的なパターンの抽出が可能となる．Linux の man3 を用いて，パターン抽出実験を行った結果，多くの対訳パターン実験を抽出でき，本手法の実現可能性を確認した．

謝辞

本研究の一部は，科研費 (基盤研究 (A)(2))「「ソフトウェア＝プログラム＋ドキュメント」の視点に基づく多言語対応大規模コーパス」によります．

参考文献

[1] E. Brill: Some Advance in Transformation-Based Part of Speech Tagging, AAAI-94 (1994).

[2] Y. Matsumoto, et al: Morphological Analysis System ChaSen version 2.2.1 Manual (2000).

[3] E. Briscoe, and J. Carroll. : Robust Accurate Statistical Annotation of General Text, LREC 2002 (2002).

[4] T. Kudoh, Y. Matsumoto: Japanese Dependency Analysis Based on Support Vector Machines, EMNLP/VLC 2000 (2000).

[5] E. Aramaki, et al: Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation, MT Summit-2001, pp.27-32 (2001).

[6] M. Ohara, et al: Automatic Extraction of Translation Patterns from Bilingual Legal Corpus, IEEE NLPKE-2003, pp.150-157 (2003).