ウイグル語―日本語機械翻訳に関するGIZA++の実験

マヒムットジャン.ママットジャン　　岡本 紘昭
朝日大学経営学研究科

GIZA++は，統計的機械翻訳で用いることを前提に作られ、翻訳モデルの部分で単語対応のアライメントを行うツールである。GIZA++はIBMモデル1～5を学習し，単語の対応関係の確率値を計算する。

今回の研究では小規模のウイグル語―日本語コーパスを作成して、実験をし、言語モデル、翻訳モデルを構築して、統計的機械翻訳でのウイグル語―日本語統計的機械翻訳システムの可能性を検討する。

1．道具の準備

１） Linuxシステムのインストール。今回使ったのはLinuxのUbuntu 10.10バージョン。

２） GIZA++ は一番新しいバージョンgiza-pp-v1.0.4を使った。

３） ウイグル語―日本語コーパス。今回は1000組程度の実験用コーパスを作成した。

４） 言語モデル作成ツールCMU (Cambridge Statistical Language Modeling Toolkit v2)を使った。

２．ウイグル語―日本語コーパスの作成

今までウイグル語―日本語コーパスがなかったため、小規模実験用コーパスを作成した。利用可能な言語資源としては

１） 中国人民日報のネットバージョン「人民網」の日本語バージョン及びウイグル語バージョンを利用した。

２） ウイグル語から日本語に訳されたウイグル語文学作品等を利用した。

３） 日本にいるウイグル人のボランティアたちが提供したウイグル語―日本語対訳例文を利用した。

これらの資料を集め、コーパス作成道具を利用してコーパスを作成した。

３．Linuxシステムをウイグル語の入出力できるように設定した。

４．言語モデルの作成

必要となるソフトをインストールしてから、日本語のテキストを利用して、日本語言語コーパスを作成する。ウイグル語―日本語コーパスの1000の日本語文章及び日本語言語コーパスを使って言語モデルを作成する。

５．翻訳モデルの作成

翻訳モデルはmkcls及びGIZA++を利用して作成する。先ず、クラスタリングツールmkclsを使ってウイグル語―日本語コーパスが含んだ単語をクラスタリングする。そして、GIZA++を使ってウイグル語、日本語単語の位置情報、統計情報を取る。

6．デコーダについて

統計的ウイグル語―日本語機械翻訳のデコーダとは、ウイグル語を受け取り、その翻訳である日本語訳を出力するシステムである。デコーダはたくさん作り出された翻訳候補の日本語文に翻訳モデルと言語モデルで確率を与え、その値が最も大きくなったものを翻訳として出力する。

7．実験結果

ウイグル語と日本語は文法的構造、語の形態的構造及び格助詞の対応など多くの面で共通の特徴があったため、自然言語処理学会の前回の年次大会で発表した「逐語訳によるウイグル語―日本語機械翻訳」の方法と同様に、統計的の機械翻訳の方法でも質の良いウイグル語―日本語機械翻訳システムが実行出来る。