2ZA-4

# WebN グラムを用いた翻訳機能付き日本語入力システム

大江究　安藤禮　松浦純樹　山中光　市村哲
東京工科大学　コンピュータサイエンス学部

## 1 はじめに

近年，電子メールや掲示板，チャットなどで多言語によるコミュニケーションの機会が増加している．母国語を翻訳するために，機械翻訳が利用されている．翻訳を行うためには翻訳サイトを訪れる必要があるが，手間がかかる．また，翻訳精度には限界があり，一度で実用的な翻訳を行うことは困難である．

本研究では，オンライン翻訳結果と折り返し翻訳結果を表示する日本語入力システムを開発した．更に，WebN グラムを利用し，漢字変換時に日本語の類語と使用頻度を表示し，より良い翻訳結果を使用者が選択できる機能を持つ．実装ツールとして IME，形態素解析ツールの茶筅，WebN グラムをデータベースにいれるために MySQL，オンライン翻訳のツールとして J-SERVER を用いた．

## 2 問題点

翻訳を行うために翻訳サイトを利用して翻訳するには手間がかかる．しかも，翻訳精度が低く，実用的な翻訳を一度で行うことは困難である．機械翻訳によって意味の変わってしまう翻訳となる例を以下に示す．

例:子牛くらいもある大きな犬がいた
翻訳文:There was a big dog around the calf.
折り返し翻訳文:子牛のまわりに大きい犬がいた

この折り返し翻訳文を見てわかるように、意味が翻訳前と異なる文章が出来る場合がある．

また，現在の英語学習では，書くことより読むことを重点的に学習しており「読めるが書けない」という状態に陥っている．図 1 は高校生の英語活動の状況である．

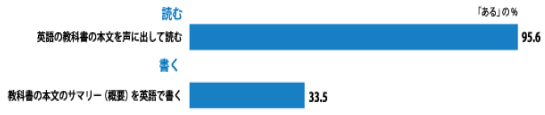


図 1:高校生の英語学習実態

英語教科書を声に出して読む学習が 95.6%に対し，教科書の概要を英文で書く学習が 33.5%となっていて，高校の英語授業が，日本人の英語を読めるが書けない原因となっているのがわかる．

## 3 提案

WebN グラムを用いて使用頻度からの英文頻度評価を表示でき，また，類語を表示，選択して自分の好きな表現を用いることができる日本語入力システムを提案する．文章入力時にオンライン翻訳システムからの翻訳結果を表示することで安心して翻訳ができるようにする．

WebN グラムとは，Google が提供している大規模言語リソースである．日本語・英語の WebN グラムがあり，日本語データは Web から抽出した 200 億文から作成した n-gram データである．出現頻度回数が日本語 20 以上，英語 40 以上の文字列のみを示す．

本システムは，翻訳された文章と折り返し翻訳，漢字変換時に類語と英語文章に対して Web 日本語，英語 N グラムから取得した頻度を表示する機能を有している．類語表示時に各語の頻度順に並び替えて表示，取得した頻度より変換された文章に視覚的に警告を示す機能を有している．

本システムの効果として，本システムを用いることにより翻訳サイトを訪れる手間が省け，表示された類語を自分で選択することによって，より良い文章表現が可能になる．また，WebN グラムから取得した頻度を確認することで，通常使われている正しい文章表現を用いることができ，異文化学習やコミュニケーションの促進をはかることができる．

A Kanji input system used Web N-gram with online translation
Kiwamu Oe, Rei Ando, Junki Matsuura, Hikaru Yamanaka, Satoshi Ichimura
School of Computer Science, Tokyo University of Technology

## 4 システム概要

図2にシステム概要図を示す.

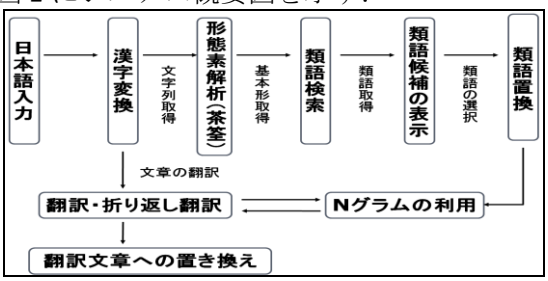


図2:システム概要

ワープロソフト等で入力された変換中の日本語未確定文字列を取得し，取得した文字列を言語グリッドに送り翻訳，折り返し翻訳を表示する．またその際に，WebNグラムを利用して翻訳文の品質を確認する．また，取得した文字列から変換中である文字列を判断し，品詞を識別，その語の基本形を取得し類語を検索して，類語の一覧を表示する．類語を選択し，入力中の文章と置換を行った際，再度WebNグラムから頻度を取得し，品質を確認することができる．また，類語の仕様頻度から上位３つを自動で選び類語の置き換え，翻訳を行うことができる．

## 5 実装

### 5.1 WebNグラムの利用

本システムではWeb日本語，英語Nグラムともに文章評価用に3-gramデータ，各単語の頻度評価のために1-gramデータを用いた．

日本語の類語置換後の文章に対してWeb日本語Nグラムによる頻度評価を行い，他の置き換え候補一覧を表示，類語候補と元の文章から置き換え候補になりうる3グラムをデータベースから取得，置き換え候補の類語表示順序を最適化するために日本語1-gramデータを用いた．

翻訳された文章に対しては，Web英語Nグラムを用いて頻度評価を行う．取得した頻度や文字列から，翻訳された文章に対し頻度が著しく少ない箇所に赤線，頻度が少ないが固有名詞があるためにそうなってしまった可能性がある場合には青線を引き，ユーザに注意を促す．

図3に，赤線が表示される例を示す．

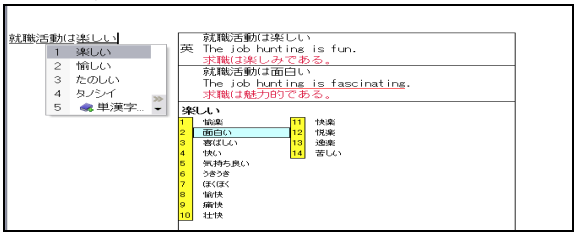


図3:英語翻訳文に対する頻度評価結果

### 5.2 候補のデータベース格納

類語の候補が見付かった場合には，その単語を再度変換した場合に少しでも早く表示できるように結果をデータベースへとキャッシュ保存する．Web検索を行う前にデータベースを参照して見つかればデータベースから取得する．

### 5.3 変換中単語の基本形取得

未確定文字列は一度茶筌にかける．茶筌とは形態素解析ツールである．形態素解析とは自然言語で書かれた文を形態素の列に分割し，品詞を見分ける作業である．変換した単語の基本形・品詞の取得する際に用いる．

## 6 まとめ

今後の課題として，キー操作を行う際に同時押しする必要があるため使い辛いという意見もありユーザインタフェースの改善も求められる．

また，データベースにWeb日本語，英語Nグラムの3-gramデータを入力する時点で，固有名詞だと思われる単語をまとめて同じ形のデータとして登録することで，頻度評価を行いやすい状態でNグラムを利用することができるようにすることができると考えられる．

## 参考文献

[1]Social IMEインターネットを用いた日本語入力システム慶應義塾大学理工学研究科　奥野 陽 萩原将文)

[2]折り返し翻訳を用いた翻訳リペアのチャットコミュニケーションへの影響（和歌山大学　宮部 真衣　吉野 孝)
情報処理学会研究報告 2009-GN-70

[3]翻訳リペアのための言い換え分自動生成手法の提案和歌山大学　宮部 真衣　吉野 孝）情報処理学会全国大会 3D-3

[4]オンライン翻訳機能を備えた日本語入力システム　AI2009-36 pp.37～42