Vol.2012-GN-83 No.11
2012/3/22

翻訳機能付きのチャットシステム

王 昕光† 高橋 佳那† 市村 哲†

近年，インターネットの普及により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．インターネット上には言語を機械翻訳するサイトや辞書サイトなどの言語処理システムが増えて来ている．
しかし，インターネット上の会話が身近になった現代では，文章中に流行語や若者言葉，ネットスラングなどの俗語が含まれることが多くなり，正確に翻訳することが困難である．また，俗語の意味を調べることや説明する手間が掛かり，その単語が本来の意味で使われているか，俗語として使われているかを確認する必要もある．
そこで本研究では，文章を翻訳する際に俗語を検出し，データベースから意味を取得してチャット上で参照できるチャットシステムを提案する．翻訳できない未知語が俗語であるかどうかは，WebN グラムを利用してチェックを行う．文章翻訳後，翻訳文だけでなく原文，折り返し翻訳文，原文に含まれた俗語とその意味を表示する．これによって，相手に送信する文章が正しく翻訳されているか確認することが出来る．また，俗語を説明したり，検索したりする手間が省ける．

Chat system with translation

Xinguang Wang † Kana Takahashi†
Satoshi Ichimura †

Recently, the opportunity of communications using multi-language increases by the spread of the Internet service. Therefore, a lot of systems such as the language dictionary sites and the machine translation sites on the Internet is also increasing.
However, today, on the Internet, much slang such as the buzzwords, the words of young people and the net slang is contained in the text. It is difficult to translate accurately. In addition, it is time-consuming to look for the meaning of the slang.
Therefore, in this research, I propose a multi-language chat system which can detect slang and check usage from the database. We can check the unknown word whether is the slang or not by using the WebN gram.
After translation, not only the translated sentence, but also original statement and statement back translation and the meaning of the slang that included in the text are displayed. We can in this way confirm whether a sentence to transmit a message to a partner is translated correctly.

1. はじめに

近年，インターネットの普及により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．例として電子メールや掲示板，チャット，SNS，インスタントメッセンジャーなどが挙げられる．それらのサービスを利用する際，利用者が相手に応じて必要な言語を扱えるとは限らない．そのため，インターネット上には言語を機械翻訳するサイトや辞書サイトなどの言語処理システムが増えて来ている．

しかし，インターネット上の会話が身近になった現代では，文章中に流行語や若者言葉，ネットスラングなどの俗語が含まれることが多くなり，正確に翻訳することが困難である．また，俗語の意味を調べることや説明する手間が掛かり，その単語が本来の意味で使われているか，俗語として使われているかを確認する必要もある．

そこで本研究では，文章を翻訳する際に俗語を検出し，データベースから意味を取得してチャット上で参照できるチャットシステムを提案する．翻訳できない未知語が俗語であるかどうかは，WebN グラムを利用してチェックを行う．文章翻訳後，翻訳文だけでなく原文，折り返し翻訳文，原文に含まれた俗語とその意味を表示する．これによって，相手に送信する文章が正しく翻訳されているか確認することが出来る．また，俗語を説明したり，検索したりする手間が省ける．

2. 従来技術の問題点

翻訳する文章に口語表現や流行語，若者言葉，ネットスラングなどの俗語が含まれる時，正しく翻訳が出来ない，また，形態素解析が正しく行われないことがある．正しく翻訳が出来ない例を以下に示す．

例:

原文: 友達にリア充が多い

翻訳文: 許多負責後方給朋友

折り返し翻訳文: 友達にリアを担当し，多くの

この折り返し翻訳から，「リア充」という俗語が正しく翻訳されず，意味が翻訳前と異なる文章になっていること，また，形態素解析が正しく行われず「リア」「充」と区切られて翻訳されていることがわかる．

† 東京工科大学 コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology

Vol.2012-GN-83 No.11
2012/3/22

3. 提案

入力文章を翻訳するときに俗語を検出し，その意味をサーバから取得して，チャットシステム上で参照できる翻訳機能付きチャットシステムを提案する．日々新しく生まれる俗語を利用可能とすることにより，文章中の俗語を正しく解析することによって誤訳を減らし翻訳精度を向上させ，より日常会話に近いやり取りが可能となる．

3.1 俗語の検出

文字列の翻訳時に形態素解析を行い，各単語をNグラムのデータベースで検索して，頻度からその単語が常用されているかどうかをチェックする．もし単語がデータベースに無い，もしくは頻度が低い場合は前後の単語と組み合わせて再確認する．

3.2 従来研究との比較

従来研究は形態素を正しく解析できない場合が多い． 図 1 は日本語形態素解析の比較を示す．

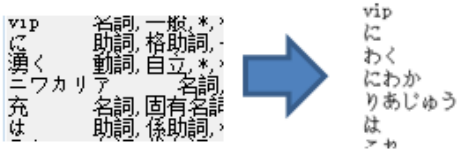

図 1 日本語形態素解析の比較

従来，翻訳機能付きのチャットシステム（QQ International[16]）は，折り返し翻訳機能を持たず，そのまま翻訳文を相手に送信するようになっている．その場合は俗語など直訳できない単語を含んでいる場合は完全に誤った翻訳をすることもある．

図 2 は従来の翻訳機能付きチャットシステム[16]の例を示す．

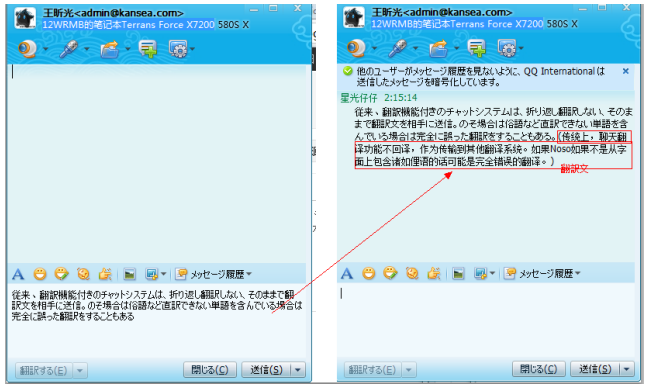

図 2 従来の翻訳機能付きチャットシステム

提案システムは，原文，翻訳文，折り返し翻訳文を表示し，原文に含んでいる俗語とその意味を表示する．相手に送信する前，翻訳文と折り返し翻訳文を確認できる．

図 3 は提案する翻訳機能付きチャットシステムの例を示す．

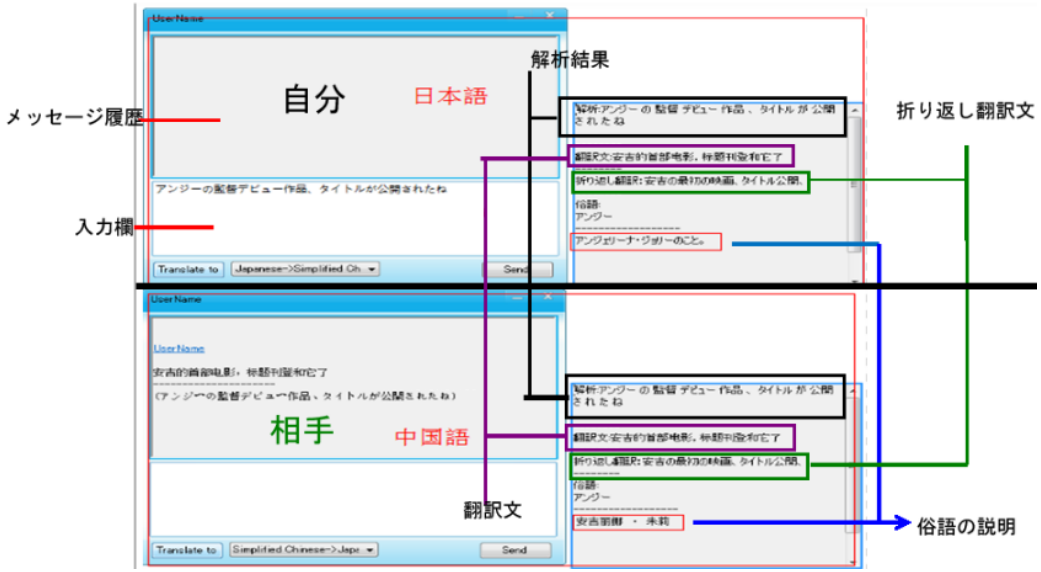

図 3 提案する翻訳機能付きチャットシステム

4. システム概要

4.1 システムの流れ

ユーザーがシステムに文字入力して，翻訳ボタンをクリックすると，文章をMecab[7]及びYahoo形態素解析APIで形態素解析する．若者言葉など含んでいる場合は解析できない文字列を取得して，俗語のデータベースで俗語と俗語の意味を取得する，それから，Microsoft翻訳API[4]で翻訳して相手に転送する．

Mecab[7]で形態素解析にかかる時間は0.02-0.05秒程度であり YahooAPI は0.5秒程度である．Mecab[7]の方が解析速度は平均 10 倍速い．よって先に Mecab[7]で形態素解析する．Mecab[7]で解析した各形態素は WebN グラムで調べ，WebN グラムにデータが少ない，または，データがない文章を含んでいたら YahooAPI で再解析する．

図 4 にこのシステムの概要図を示す．

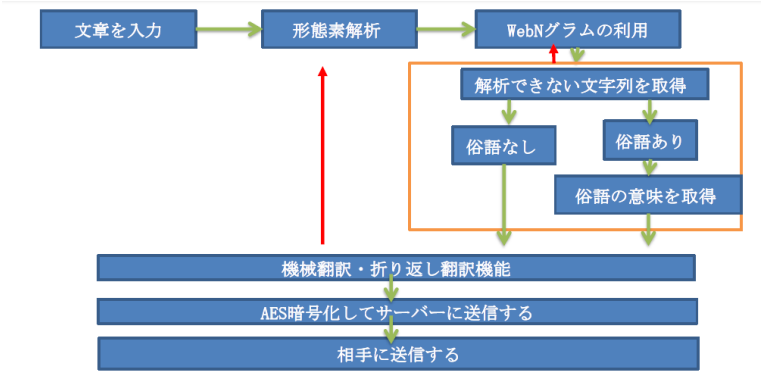

図 4　システムの概要図

4.1 解析できない文字列の取得

文章に俗語や未知語などが含まれる時には Mecab[7]で正しく解析できない場合がある．

例として，「逆に，よく vip に湧くニワカリア充はこれがないから無駄にスレを汚す」を Mecab[7]で解析したら図 5 のように，俗語「ニワカ」と「リア充」は「ニワカリア」と「充」に分かれて解析された．このような時，解析した各単語を WebN グラムのデータベースで検索して，データベースに載っていない単語，または，頻度低い文字列があることを検出できる．

```
逆　名詞,一般,*,*,*,*,逆,ギャク,ギャク
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
、　記号,読点,*,*,*,*,、,、,、
よく　副詞,一般,*,*,*,*,よく,ヨク,ヨク
vip　名詞,一般,*,*,*,*,*
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
湧く　動詞,自立,*,*,五段・カ行イ音便,基本形,湧く,ワク,ワク
ニワカリア　名詞,一般,*,*,*,*,*
充　名詞,固有名詞,人名,名,*,*,充,タカシ,タカシ
は　助詞,係助詞,*,*,*,*,は,ハ,ワ
これ　名詞,代名詞,一般,*,*,*,これ,コレ,コレ
が　助詞,格助詞,一般,*,*,*,が,ガ,ガ
ない　形容詞,自立,*,*,形容詞・アウオ段,基本形,ない,ナイ,ナイ
から　助詞,接続助詞,*,*,*,*,から,カラ,カラ
無駄　名詞,形容動詞語幹,*,*,*,*,無駄,ムダ,ムダ
に　助詞,副詞化,*,*,*,*,に,ニ,ニ
スレ　名詞,一般,*,*,*,*,*
を　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
汚す　動詞,自立,*,*,五段・サ行,基本形,汚す,ケガス,ケガス
EOS
```

図 5　Mecab による形態素解析

4.2 形態素解析結果の直し

Mecab[7]によってうまく解析できない文字列を取得した場合は，その単語とその単語前後の単語の組を Yahoo の形態素解析 API で再度解析する．Yahoo の形態素解析 API による解析結果を図 6 に示す．


```
<ResultSet xmlns:xsi="http://www.w3.org/2001/XI
 http://jlp.yahooapis.jp/MAService/V1/parseResp
  <ma_result>
    <total_count>3</total_count>
    <filtered_count>3</filtered_count>
    <word_list>
      <word>
        <surface>湧く</surface>
        <reading>わく</reading>
        <pos>動詞</pos>
      </word>
      <word>
        <surface>ニワカ</surface>
        <reading>ニワカ</reading>
        <pos>名詞</pos>
      </word>
      <word>
        <surface>リア充</surface>
        <reading>りあじゅう</reading>
        <pos>名詞</pos>
      </word>
    </word_list>
  </ma_result>
```


図 6　API による形態素解析 1

しかし，Yahoo の形態素解析 API も正しく解析できない場合もある．

例として，「マルモリ」と Yahoo の形態素解析 API で解析したら図 7 のように，俗語「マルモリ」は「マル」と「モリ」に分かれて解析された．


```
<ResultSet xmlns:xsi="http://www.w3.org/2
 http://jlp.yahooapis.jp/MAService/V1/pars
  <ma_result>
    <total_count>2</total_count>
    <filtered_count>2</filtered_count>
    <word_list>
      <word>
        <surface>マル</surface>
        <reading>まる</reading>
        <pos>名詞</pos>
      </word>
      <word>
        <surface>モリ</surface>
        <reading>もり</reading>
        <pos>名詞</pos>
      </word>
    </word_list>
  </ma_result>
</ResultSet>
```


図 7　API による形態素解析 2

Yahoo API で解析した単語を正しい解析結果と判断出来ない場合は，解析結果の各単語を俗語のデータベースで検索して，その単語を含んでいる場合は，その単語とその単語前後の組は俗語と判断出来る．例として，「マル」「モリ」は「マルモリ」に変換する．

図 8 に俗語のデータベースで「マル」を検索した結果を示す．

```
Array
(
    [0] => Array
        (
            [id] => 2813
            [word] => マルモリ
            [hiragana] => まるもり
            [lang] => japanese
        )

    [1] => Array
        (
            [id] => 2812
            [word] => マル政
            [hiragana] => まるせい
            [lang] => japanese
        )

    [2] => Array
        (
            [id] => 2811
            [word] => マルガリーマン
            [hiragana] => まるがりーまん
            [lang] => japanese
        )

)
```

図 8 「マル」を俗語のデータベースで検索した結果

5. 要素技術

5.1 形態素解析

(1) 日本語

Mecab[7]はオープンソースの形態素解析エンジンで，自然言語で書かれた日本語文を形態素の列に分割し，品詞を見分けるソフトウェアである．変換した単語の基本形・品詞の取得，名詞に付属する修飾語を探す際に使用した．

図 9 に「今日はいい天気です」という一文を Mecab[7]で形態素解析にかけた結果を示す．

```
日本語：今日はいい天気です
    今日        名詞,副詞可能,*,*,*,*,今日,キョウ,キョー
    は          助詞,係助詞,*,*,*,*,は,ハ,ワ
    いい        形容詞,自立,*,*,形容詞・イイ,基本形,いい,イイ,イイ
    天気        名詞,一般,*,*,*,*,天気,テンキ,テンキ
    です        助動詞,*,*,*,特殊・デス,基本形,です,デス,デス
```

図 9 MeCab による形態素解析

Mecab でうまく解析できない場合，「Yahoo 日本語形態素解析 API」を利用して再解析する．

表 1 に本研究が使っている日本語形態素解析 API のリクエストパラメータを示す．

表 1 日本語形態素解析 API のリクエストパラメータ

パラメータ	値	説明
appid	String	アプリケーション ID.
sentence	String	解析対象のテキストです.

(2) 中国語

ICTCLAS[15] は Institute of Computing Technology， Chinese Lexical Analysis System の略で，中国科学院計算技術研究所が開発した中国語形態素解析ツールである．

図 10 は ICTCLAS[15]で形態素解析した例を示す．

```
中国語：今天天气不错
    今天        /t
    天气        /n
    不错        /a
```

図 10 ICTCLAS による形態素解析

図中，/t は時間詞，/n は名詞，/a は形容詞[17]である．

5.2 WebN-グラム

WebN-グラムは web 上に存在する単語や語句(n-gram)の出現頻度をまとめた大規模言語リソースである．Google が作成したものが存在する．この日本語 N グラム，英語 N グラム及び中国語 N-グラムを LDC(Linguistic Data Consortium)から購入して用いた．ここでの n-gram 言語モデルとは，直前の(n-1)個の単語を見て，次の単語を予測するモデルのことである．

図 11 は日本語，英語，中国語の 3-gram の例を示す．

日本語Nグラム		中国語Nグラム		英語Nグラム	
情報 連絡 会	1723	歌曲 作为 编写	46	citation in scientific	85
情報 連絡 会議	590	歌曲 作为 美国	70	citation in section	104
情報 連絡 体制	1795	歌曲 作为 联络	45	citation in some	58
情報 連絡 係	25	歌曲 作为 背景	2089	citation in subsection	42
情報 連絡 先	831	歌曲 作为 自己	793	citation in such	100
情報 連絡 協議	191	歌曲 作为 舞蹈	136	citation in support	218
情報 連絡 及び	104	歌曲 作为 艺术	77	citation in text	791950

図 11 3-gram の例

図のように，３つの単語に分解されたものとその出現頻度が記されている．

この WebN-グラムデータを MySQL データベースに格納することで高速な検索が可能となるようにした．MySQL データベースには中国語と英語の 5-グラムまでのデータ及び日本語の 7-グラムまでのデータを格納した．

本研究では，この WebN-グラムの並びと頻度を利用して英文全体について頻度評価を行ったり，日本語の文章を正規化，俗語頻度の判定を行ったりするために用いている．

図 12 は Web 日本語 N グラムにおける『ネタにマジレスウケる』を解析したデータ例を示す．

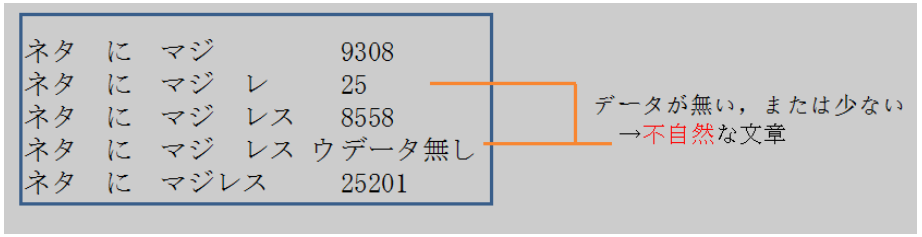

図 12 『ネタにマジレスウケる』を解析した例

例のように「ネタ　に　マジ　レ」の頻度は 25 で少なく，「ネタ　に　マジ　レス　ウ」はデータがない．よって「ネタ　に　マジ　レ」と「ネタ　に　マジ　レス　ウ」は不自然な文章と判断できる．

5.3 俗語データベース

俗語の意味を取得し，各俗語辞書サイトから俗語，読み仮名，説明を取得して俗語データベースに登録した．入力文を形態素解析する際に，基本形や読み仮名を取得し，それを用いて検索する．

図 13 に，実装に用いた俗語辞書サイトを示す．

図 13　俗語辞書サイト

現在は日本語俗語の取得先として「げんごや」俗語サイトから約 3,000 俗語を取得した．英語俗語の取得先として「Slang」俗語サイトから約 1000 俗語を取得した．中国語俗語の取得先として「百度検索ランキング」新用語サイトから約 100 俗語を取得した．

5.4 Microsoft 翻訳 API

Microsoft より提供されている翻訳および言語 Web サービスである．現在 37 種類の言語に対応している．AJAX/SOAP/HTTP の各インターフェースで利用可能．本研究ではこの API を利用して機械翻訳を行っている．

表 2 に本研究が使っている Microsoft 翻訳 API[4]のリクエストパラメータを示す．

表 2　Microsoft 翻訳 API のリクエストパラメータ

パラメータ	説明
appId	A string containing the Bing AppID.
Text	A string representing the text to translate.
From	A string representing the language code of the translation text.
To	A string representing the language code to translate the text into.
maxTranslations	An int representing the maximum number of translations to return.

6. まとめ

入力文章を翻訳するときに俗語を検出し，その意味をサーバから取得して，チャットシステム上で参照できる翻訳機能付きチャットシステムを提案した．文字列の翻訳時に形態素解析を行い，各単語を N グラムのデータベースで検索して，頻度からその単語が常用されているかどうかをチェックする．もし単語がデータベースに無い，もしくは頻度が低い場合は前後の単語と組み合わせて再確認する．相手に送信する前，翻訳文と折り返し翻訳文を確認できる．

文章翻訳後，翻訳文だけでなく原文，折り返し翻訳文，原文に含まれた俗語とその意味を表示する．これによって，相手に送信する文章が正しく翻訳されているか確認することが出来る．また，俗語を説明したり．検索したりする手間が省ける．

提案したシステムにより，Web サイトを利用しながらの会話に比べ手間は少なくなり，早く翻訳結果を得られるようになった．

今後の課題を示す.

１．　動作速度の高速化

２．　俗語辞書の充実

１について，データ送受信の付加を減らすため，通信方式を見直す必要がある．また，データベース検索の高速化を行うことが必要と考えられる．翻訳に時間がかかる場合は読込み中であることが分かるようにする．

2 について，英語と中国語の俗語データは少ないである．俗語をサイトから自動取得する．また，ユーザーが登録できるようにすると考えられる．

参考文献

1) Yahoo!類語辞書
http://dic.yahoo.co.jp/
2) エキサイト翻訳
http://www.excite.co.jp/world/
3) Yahoo!翻訳
http://honyaku.yahoo.co.jp/
4) Microsoft Translator API
http://www.microsofttranslator.com/dev/
5) 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
6) 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
情報処理学会研究報告　2009-GN-70　宮部真衣 吉野孝
7) MeCab　(形態素解析ツール)
http://mecab.sourceforge.net/
8) iconv　（文字コード変換）
http://www.gnu.org/software/libiconv/
9) げんごや.com　ジャパ造語
http://gengoya.com/japazougo
10) Slang
http://www.chatslang.com/list/o
11) Internet Slang Dictionary & Translator
http://www.noslang.com/search.php
12) 富士見中学校多言語対話システム
http://langrid.org/playground/fujimi-translation.html
13) 翻訳リペアのための言い換え分自動生成手法の提案
情報処理学会第 72 回全国大会 3D-3 和歌山大学　宮部 真衣　吉野 孝
14) 文法に基づいて機械翻訳システム
南京大学　南京 21009 陳家骏　戴新宇　尹存燕　王启祥
15) ICTCLAS （形態素解析ツール）
http://ictclas.org/
16) QQ International （翻訳機能付きのチャットシステム）
http://www.imqq.com/
17) ICTCLAS 品詞標示
http:// www.nlp.org.cn/docs/docredirect.php?doc_id=993