社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－171 (15)
2006／1／13

WWW 上のテキスト情報を利用した翻訳品質評価法の検討

宮下広平†*, 安田圭志‡, 山本誠一‡†, 柳田益造†

†同志社大学
‡ATR 音声言語コミュニケーション研究所

日英方向の機械翻訳による訳文の正解訳と誤訳を分類する手法として, WWW 上に存在する膨大な英文の用例に着目し, WWW 上で翻訳文が検索ヒットするか否かにより正解訳と誤訳を分類する. 誤訳を訳文の流暢さ (Fluency) と適切さ (Adequacy) により 3 つのクラスに分類し, 正解訳と各誤訳のクラスの検索ヒット率が, 文中に含まれる単語数に応じてどのように変化するかについて検討している. 本手法を用いることで, 対象とする分野の会話文に適合した機械翻訳を用いれば, かなりの割合で正解訳と誤訳を分類可能であることを示し, 更に本手法の幾つかの課題を指摘すると共に, 単語数に応じた閾値の設定などの課題の対処法を述べる. 最後に, 市販されている機械翻訳システムによる翻訳文に対して本手法の適用可能性について検討している.

Quality Evaluation Method of Machine Translated Sentences by Comparing Text Retrieved from WWW

Kohei Miyashita†*, Keiji Yasuda‡, Seiichi Yamamoto‡†, and Masuzo Yanagida†

†Doshisha University
‡ATR Spoken Language Translation Research Labs.

This paper proposes a method for evaluating quality of machine translated sentences by comparing them with text retrieved from WWW. We conducted an experiment on investigating how many sentences exactly matched to correct translations and incorrect translations are retrieved from WWW using translations from several translation systems. Our approach to classify correct and incorrect translations is effective for medium length sentences consisting of five to nine words.

1 はじめに

経済のグローバル化の進展に伴い, 英語は国際共通語としての役割を担う傾向が加速し, 特定の分野だけではなく様々な分野で英語を使って意見を述べたり情報を発信することが重要になってきている.

英語学習者の分野の広がりに伴い, 学習者のニーズに合わせて対象分野を限定し, その分野特有の学習内容を提示することにより学習効果を高める学習法である ESP(English for Specific Purposes)が注目されている. 更に ESP に於いても, 対象分野での英語での発信能力を向上させることが重要な課題の 1 つとなりつつある.

音声による英語での発信能力を向上させる手段の 1 つに, 「英語対話システム」を用いた学習が考えられる. 筆者らは, 英語学習者がコンピュータと様々な課題について英語で対話を行う際に, 学習者の英語の能力測定を行い, 能力に応じて課題を変更することにより, 発信能力の向上を支援する「英語対話システム」の開発を進めている. 「英語対話システム」において, 目標分野毎に対話シナリオを作成することは, 多くの人的資源, 費用, 時間を要するために, 如何にして開発を効率的に行うかが課題になる. この課題に対して, 既存の特定分野の英語対話システムを対象分野に移植する方法や日本語対話システムをベースに英語対話システムを開発することとし, その際の効率的な開発手段として, 機械翻訳などの自然言語処理技術を使用することが考えられる.

現在, World Wide Web(WWW)上の翻訳サービスの普及

* この研究の一部は ATR 音声言語コミュニケーション研究所で行われた.

や，大規模な対訳コーパスの開発を基盤とした統計翻訳システム，用例翻訳システムによる高品質な翻訳の実現により，機械翻訳が広く一般に使われるようになった．しかし，機械翻訳による翻訳品質の精度が向上したとはいえ，誤訳文を生成してしまうケースがある．このため，翻訳された文が適切な表現であるかどうかを検証することが重要となるが，全ての訳文の検証を人手で行うことは，機械翻訳による効率化の効果を損なうこととなる．

本研究では，翻訳文の品質を評価する手法として，WWW 上に存在する文書群の利用を検討する．WWW 上のテキスト情報を利用して英文の品質を識別する手法は，隅田らにより，英語の多肢選択課題の自動作成手法について，その妥当性を検証するための手段として提案されている[1]．隅田らの手法では，英文の多肢選択課題で棄却候補を自動作成する際に，WWW 上に存在するテキストは適切な英文であるとみなし，WWW 上に存在するテキストと同一の表現の課題は多肢選択課題の棄却課題としては適切でないとして排除する．

同様に，機械翻訳システムからの翻訳文を WWW 上で検索することができれば，その翻訳文は正しい表現と見なせる確率が高いと推測できる．正しい翻訳と見なせる文は適切な表現であるか否かの検証の対象外とすることにより，訳文の品質の検証過程を効率化できる．

本稿では，機械翻訳による複数の翻訳文の中から，翻訳品質の高い翻訳文のみを分類することを目的とし，WWW 上のテキストとマッチングを行うことにより翻訳品質を評価する手法について検討した結果を報告する．

第 2 章では，提案手法について説明し，実験結果を示すと共に，実験により示された提案手法の課題について説明する．第 3 章では，これらの課題に関する対処法を検討する．第 4 章では，市販されている機械翻訳を用いる場合の問題点の検討結果について述べる．第 5 章では，まとめと今後の検討課題について述べる．

2 提案手法と検索ヒット率

既に述べたように，WWW 上には様々な英文テキストを見つけることができ，翻訳文と完全一致する文を WWW 上で検索することができれば，その翻訳文は正しいと見なせる確率が高く，この文を正解訳と誤訳の評価対象から除くことにより効率化が可能と推測できる．

本稿では，本推測の妥当性の検証を行うために，まず想定する英文が WWW 上に検出される割合を調査する．対象とする英文としては，ATR で収集された，旅行会話に関する大規模な日英対訳コーパスである BTEC（Basic Travel Expression Corpus）[2]の中からランダムに選択された約 2,300 文の英文を用いた．

2.1 適切な英文の検索ヒット率の検証

検索エンジン Google[3]により，評価対象文の英文と完全一致する文を WWW から検索した際の検索ヒット率を図 1 に示す．図 1 の横軸は英文に含まれる単語数を，縦軸は単語数毎の検索ヒット率を示す．尚，検索ヒット率とは，全ての翻訳文中の 1 件でも完全一致する文が存在する翻訳文の割合を示している．

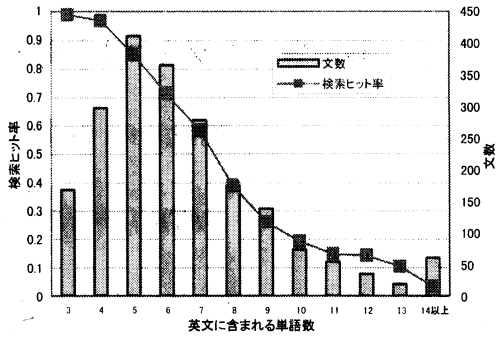

図 1：英文に含まれる単語数毎の検索ヒット率

評価対象文全体としての検索ヒット率は 68%であった．図 1 では省略しているが，1 文に含まれる単語数が 1 または 2 である文は，全体の 6.5%あり，検索ヒット率は 100%であった．図 1 から，英文に含まれる単語数が多くなるに伴い，検索ヒット率が低下していることがわかる．また，BTEC の英文の平均単語数は 6.2 であり，単語数が 6 のとき，検索ヒット率は 71%，単語数が 7 のとき，検索ヒット率は 58%と，平均単語数周辺で検索ヒット率が 50%を超えた．一方で，単語数が 10 以上となると，検索ヒット率が 20%以下まで低下した．

以上から，英語として適切な表現と考えられる BTEC の英文に関しては，1 文に含まれる単語数が極端に多くない場合は，かなりの割合で検索ヒットし，WWW上に完全一致する文が存在し，英文の質を検証するためのリソースとして，WWW上のテキストを検索する手法は有効に機能すると考えられる．

適切な英語表現の文では，WWW 上に完全一致する文がかなりの割合で存在するが，機械翻訳による誤訳も検索ヒットすることが予測される．次節では，実際に機械翻訳による翻訳結果を用い，それらを正解訳と誤訳に分類して，検索ヒット率の検証を行う．

2.2 機械翻訳による訳文の検索ヒットの検証

機械翻訳による翻訳文の検索ヒットの検証を行うにおいて，BTEC が対象とする旅行会話に関する文に適合していると考えられる 4 種類の異なる機械翻訳方式による翻訳文を用いた．具体的な翻訳文は，BTEC を対象とした機械翻訳に関するワークショップである IWSLT2004 評価キャンペーン[4]における言語資源と参加した翻訳システムの翻訳結果を利用した．このキャンペーンは，旅行会話に関する BTEC を用い，参加システムの翻訳結果を評価するものである．翻訳文の総数は，評価キャンペーンに使用された BTEC のテスト文 500 文に対する翻訳結果，計 2,000 文である．なお，図 1 に示した結果を得るために使用した約 2,300 文の英文は，その一部として本ワークショップで使用された 500 文の対訳を含んでいる．

翻訳文の WWW 上での検索ヒットの検証手順は以下の

通りである．

1. 日英バイリンガル3名により，翻訳文ごとの翻訳品質を Fluency（流暢さ）と Adequacy（適切さ）の観点からの主観評価を，表1の評価基準に基づき行う．3 人の評価値のメディアン値を評価値として採用する．翻訳品質の主観評価を用いて，翻訳文を正解訳（Fluency, Adequacy 共に4以上）と3つの誤訳クラスに分類．誤訳クラスとしては，誤訳クラスⅠ（Fluency, Adequacy 共に3以下），誤訳クラスⅡ（Fluency は3以下，Adequacy は4以上），誤訳クラスⅢ（Fluency は4以上，Adequacy は3以下）に分類．
2. 翻訳文を検索フレーズとして，翻訳文と完全一致する文の検索ヒット文数（WWW 上に完全一致する文が検索される場合の入力された翻訳文の数）を求める．

表1：翻訳品質の主観評価

Fluency *1	
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

Adequacy *2	
5	All Information
4	Most Information
3	Much Information
2	Little Information
1	None

*1 Fluency(流暢さ)：英語としてどのくらい滑らかか
*2 Adequacy(適切さ)：元の文の内容がどの程度正確に反映されているか

図2に翻訳文に含まれる単語数毎の正解訳とクラスⅠ，Ⅱ，Ⅲに分類される誤訳の文数を示す．なお，以下の説明の簡単化のために，各翻訳文に含まれる単語数に応じて，表2に示すように翻訳文を，単語数が5未満（単語数クラスⅠ），単語数が5〜9（単語数クラスⅡ），単語数が10以上（単語数クラスⅢ）に分類する．

表2：翻訳文の単語数に応じたクラス分類

クラス	単語数	占める割合
単語数クラスⅠ	5未満	27%
単語数クラスⅡ	5〜9	59%
単語数クラスⅢ	10以上	14%

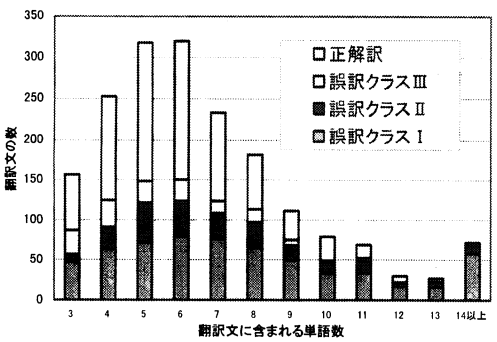

図2：単語数毎の正解訳と誤訳の文数

翻訳文と完全一致する文を，Google を用いて検索した場合の，検索ヒットした翻訳文の文数を図3に示す．図3の横軸は翻訳文に含まれる単語数を，縦軸は検索ヒットした翻訳文の文数を表している．

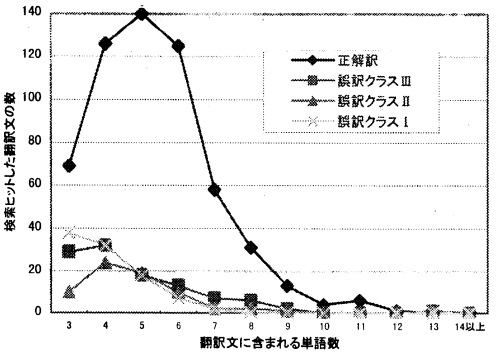

図3：単語数と検索ヒットした翻訳文の数の関連

図3から，翻訳文に含まれる単語数が単語数クラスⅡの場合は，検索ヒットした文数に顕著な差を確認できた．つまり，検索ヒットするか否かによる正解訳と誤訳の分類は，対象とする分野の会話文に適合した機械翻訳システムを用いた場合，かなりの割合で可能であると考えられる．

3 提案手法の課題と改善

WWW 上で検索ヒットするか否かで，正解訳を分類する手法を 2 章で提案し，単語数がある程度以上の文の場合，かなりの割合で分類可能であることを示した．しかし，正解訳と誤訳の分類を可能とするためには，まだ多くの課題が残っている．以下，幾つかの課題と対処法について述べる．

図4に単語数と検索ヒット率の関係を示す．図4から明らかなように，比較的出現する文数の多い単語数で7〜10の中程度の長さの文の場合，検索ヒット率が低下する．更に，同じく中程度の長さの文について誤訳クラスⅢ（Fluency は4以上，Adequacy は3以下）の誤訳の割合が高い問題などが存在する．

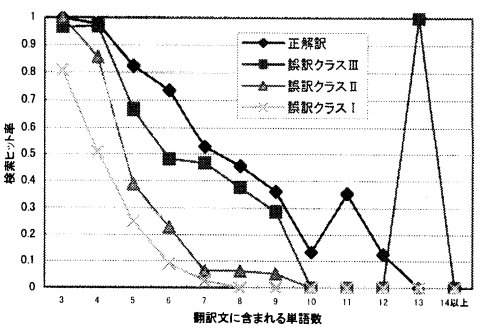

図4：単語数に対する検索ヒット率

3.1 検索ヒット数の閾値の設定

Fluency が高い誤訳文に関しては，元の日本文の意味を適切に伝えているかという点には問題があるが，英文の表現としては適切であるので，WWW 上で検索ヒットしやすいと想定される．一方，Adequacy が低い翻訳文が検索ヒットする場合，その検索ヒット数は，正解訳が検索ヒットする場合と比較して少ないと推測できる．そこで，入力された 1 つの翻訳文に対し，WWW 上で検索ヒットした文の数である検索ヒット数に，ある値を超えない検索ヒット数は検索ヒットしたと認めないという閾値を設けることにより，翻訳文を検索ヒットするか否かのみで分類するより精度が上がると考えられる．

正解訳とクラス I 及びクラス II に分類される誤訳の検索ヒット数を，単語数毎に分類した結果を図 5〜図 7 に示す．

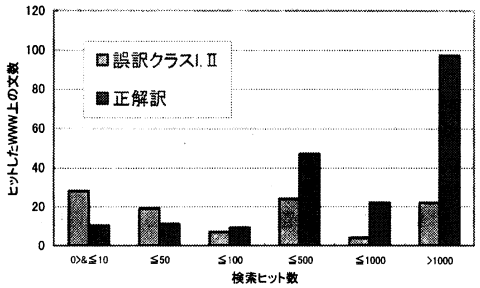

図 5：単語数クラス I の検索ヒット数の比較

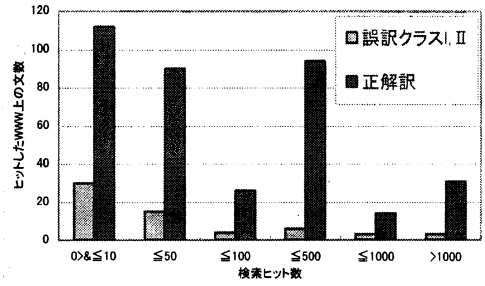

図 6：単語数クラス II の検索ヒット数の比較

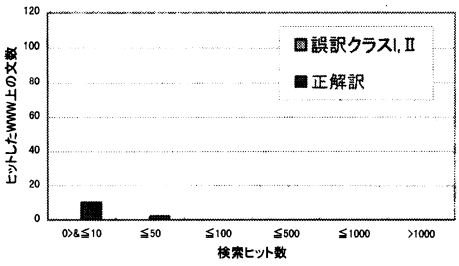

図 7：単語数クラス III の検索ヒット数の比較

図 5 に示されるように，単語数クラス I のときでは，Adequacy が低い文でも，正解訳と同等の検索ヒット数を持つ文も少なくない．一方，図 6 に示す単語数クラス II の場合では，検索ヒット数が 100 を超える Fluency の低い翻訳文はほとんど出現しないことがわかる．また，図 7 から明らかなように，単語数クラス III の場合は，そもそも検索ヒットする Fluency が低い翻訳文は，ほぼ出現しない．以上から，翻訳文に含まれる単語数が少ない場合を除けば，Fluency が低い翻訳文と正解訳に関して，検索ヒット数に顕著な差が出る．

次に，翻訳文の閾値の設定により，検索ヒットした文の中に含まれる正解訳の割合の変化を示す．図 8 に，ある閾値を超えないヒット文数の場合は，検索ヒットと認めないという閾値を設定した場合の結果を示す．検索ヒット数の閾値を設定するにあたり，検索ヒット数を以下の式により正規化する．なお，識別力の弱い単語数 3，4 の翻訳文は除いてある．尚，図 8 の縦軸の第 1 軸は正解訳の検索ヒット率を，縦軸の第 2 軸は検索ヒットした訳文中の正解訳の割合を示している．

$$\text{正規化検索ヒット率} = \frac{\text{検索ヒット数}}{\text{同単語数の翻訳文の検索ヒット数の平均}}$$

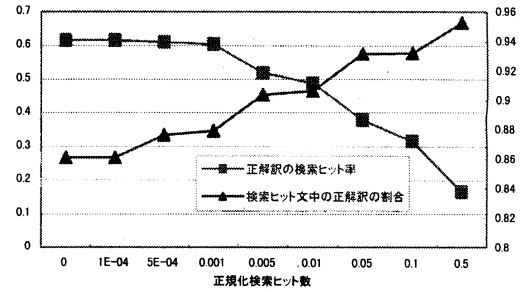

図 8：閾値を変化させた場合の正解訳のヒット率と正解訳の占める割合の関係

図 8 で示されているように，Adequacy の低い誤訳文をほぼ削除するまで閾値を上げていくと，正解訳の検索ヒット率も大きく下げてしまう．しかし，検索ヒット中の正解訳の割合を 9 割まで上げるに必要な閾値では，正解訳の検索ヒット率が約 1 割の低下で済む．以上から，検索ヒット数の閾値を設定することにより，Adequacy の低い誤訳文を一定の割合で削除可能であると考えられる．

3.2 単語クラスの導入などによるヒット率の向上

正解訳が WWW 上で検索ヒットしない要因として，以下のようなことが考えられる．

1. 訳文を構成する単語に，数詞や固有名詞，頻度を示す副詞などのクラス内で置換可能な語が含まれている．
2. 訳文に，"please"などの会話特有な挿入表現が含まれる．
3. 訳文に，"I'd"などの短縮形が含まれている．

表3に，WWW上で正解訳が検索ヒットしないと考えられる要因別の割合を示す．

表3：ヒットしない正解訳(239文)の要因別の割合

短縮形を含む文	69文/239文(29%)
会話特有な挿入表現を含む文	49文/239文(21%)
数詞を含む文	39文/239文(17%)
固有名詞を含む文	19文/239文(8%)

以下，これらの要因に対する検索ヒット率の改善手法について述べる．

① 要因1に対しては，汎用性の高い品詞に対して，同じ品詞内でクラスを作り，そのクラス内の単語が訳文に含まれた場合，その単語を，属するクラスに置き換えて検索する．

② 要因2に対しては，全体としての意味に影響しないと考えられる"please"，"hi"，"hello"などの会話特有な挿入表現を削除する．

③ 要因3に対しては，短縮形を，短縮形と短縮しない形の双方を用いて検索する．

上記の修正を行うことによる検索ヒット数の向上の結果を表4に示す．本検証での数詞をクラスに置換する方法は以下の例のような簡便な手法を用いた．

(例) *I'd like two brandies.* ⇒ *I'd like (one OR two OR ··· ten) brandies.*

上記の方法で検索を行うと，一文に数詞が2つ以上含まれた場合，Googleの検索語数の上限を超えてしまう．本検証では，このような場合は置換を行っていない．

表4：訳文修正で検索ヒット可能になった文の割合

短縮形の修正	33文/69文(48%)
会話特有な挿入表現除去	21文/49文(43%)
数詞をクラスに置換	14文/39文(36%)

翻訳文に対して，上記の改善手法①～③を全て実施した結果を図9に示す．図9には，正解訳の訳文修正前のヒット率と修正後のヒット率を示す．併せて，誤訳クラスⅠ及びⅡの修正前と修正後のヒット率を示す．

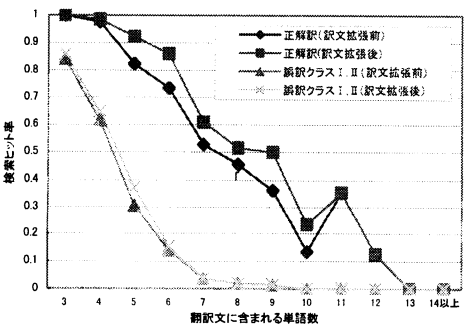

図9：検索ヒット率の改善手法によるヒット率の改善

単語数クラスⅡ（単語数5～9）に関して，正解訳は訳文修正後にヒット率が約10%程度向上していることがわかる．一方で，誤訳クラスⅠ及びⅡは，ほぼヒット率は変化していない．以上から，本手法は，単語数クラスⅡで，Fluencyの低い誤訳のヒット率を変えずに，正解訳の検索ヒット率の向上を可能にしている．

4 市販の機械翻訳を用いた検討

英語対話システムの対象分野毎の対話シナリオの開発の際に，市販されている機械翻訳を使用することが考えられる．このため，市販の機械翻訳を用いたときに，正解訳と誤訳で検索ヒット率に顕著な差が出るかどうかが問題になってくる．そこで，市販されている機械翻訳を用いて，訳文の検索ヒット率を調査した．翻訳対象文は2.2節で用いたテスト文と同様のBTECの日本文500文を用いた．また，市販されている2種類の機械翻訳を用いて対象文を翻訳した．尚，使用した機械翻訳に対して，弥勒↔Maitreyaなど未登録語に関しては，一部辞書項目などの追加を行っている．翻訳文の検索ヒットの調査手順は，日英バイリンガルが1名であることを除いて，2章の評価手法と全く同じ評価手法である．調査結果を図10，図11に示す．

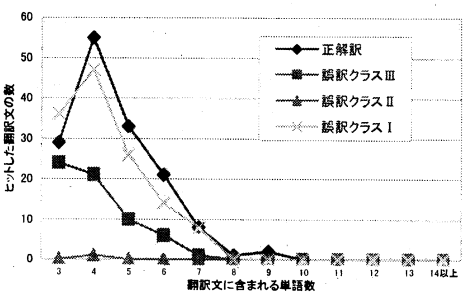

図10：会話文の訳文の調査結果1

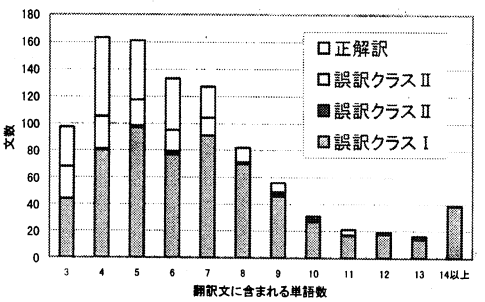

図11：会話文の訳文の調査結果2

図11で示されているように，Fluency，Adequacyともに低い翻訳文が他よりも多く，全体の66%であった．その結果，図10では，Fluency，Adequacyともに低い翻訳文の検索ヒットした文数が目立つ．市販の機械翻訳による訳文は，2.2節における4種類の機械翻訳による訳文と比較して，誤訳文が多くなっている．このため，市販の機械翻

訳の翻訳結果を使用するには，未知語登録以外に会話特有の表現などへの対応が必要と考える．

5 結論と今後の課題

本稿では，「英語対話システム」の開発の効率化のために，機械翻訳を用いることを想定し，機械翻訳の正解訳と誤訳を分類する手法として，WWW 上で翻訳文が検索ヒットするか否かによって正解訳と誤訳を分類する手法について検討した．WWW 上のテキストの検索には Google を使用した．本手法では，翻訳の対象とする分野の会話文に適合した機械翻訳システムを用いれば，BTEC のような会話文の場合，会話表現の 60%を占める単語数が 5〜9 の翻訳文については，正解訳と誤訳では，検索ヒットした訳文数に顕著な差があることを確認した．

一方で，Fluency が低い文でも一定の割合でヒットする問題などがあった．これらの改善手法として，訳文の検索ヒット数に閾値を設ける手法及び単語クラスの導入により，全体の60%の出現割合を示す単語数5〜9の翻訳文で，Fluency の低い訳文のヒット率を下げ，正解訳のヒット率を上げることが示された．

今後は以下のような課題に関して検討を行う．

① Google を使用する現在の検索手法では，検索する際に，文に含まれるピリオドは無視される．その結果，文としてではなく，部分文として検索される場合がある．これにより，"how would you like."や"may I'd like."などの Fluency の低い訳文もヒットする．この改善手法として，テキスト文そのものを獲得し，テキスト文の中から訳文を探し，訳文が部分文ではなく文として一致しているかを検証する．

② 市販ソフトの機械翻訳により会話文などの口語表現を翻訳した場合，誤訳率が高くなる．このため，市販の機械翻訳の翻訳結果を使用するには，未知語登録以外に会話特有の表現などへの対応を検討する．

③ Adequacy が低く Fluency が高い誤訳である誤訳クラスⅢの翻訳文の識別には，WWW 上のテキスト検索のみでは解決が困難であると考えられる．このため，正解訳の識別には，関連研究でなされている 2 言語の対訳文の句アラインメント法[5]や，IBM モデル I などの翻訳モデルの利用[6]などを検討する．

謝辞

本研究を進めるにあたり，有意義なコメントを頂いた ATR 音声言語コミュニケーション研究所，隅田英一郎主幹研究員，ルパージュ・イブ主任研究員に感謝致します．

本研究は，科学研究費補助金（基盤研究B）（課題番号16300048）による助成研究の一部である．

参考文献

(1) E. Sumita, F. Sugaya, S. Yamamoto：“Measuring non-native speaker's profiency of Englsih by using a test with automatically-generated fill-in-the-blank questions”, Proc. 2nd Workshop on Building Applications using NLP, pp. 61-68 (2005).

(2) T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, S. Yamamoto：“Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the rea world”, Proc. LREC-02, pp.147-152 (2202).

(3) “http://www.google.com/apis/index.html”

(4) Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, J. Tsujii.:“2004. Overview of the IWSLT04 evaluation campaign”, Proc. IWSLT2004, pp.1-12 (21004).

(5) 今村賢治：“構文解析と融合した階層的句アライメント”，自然言語処理，言語処理学会，Vol.9, No.5, pp.23-42, 2002.

(6) 土居誉生，隅田英一郎：“単語翻訳モデル駆動型の翻訳後編集”，情報処理学会　研究報告 2005-NL-169, pp. 13-18, 2005.