2008－NL－184 (14)
2008／3／28

統計的機械翻訳における翻訳誤り原因自動同定手法の提案

林克彦† 山本誠一†
†同志社大学

概要

本稿では統計的機械翻訳システムにおいて翻訳誤りを引き起こす原因として言語モデル，翻訳モデル，デコーダの3つを考え，個々の翻訳事例でそれら3つのうちから翻訳誤りの原因となった要素を自動同定する手法を提案する．提案手法では入力文に対する翻訳結果と参照訳（リファレンス）のモデル尤度を決定木から比較することで原因同定を行う．また，提案手法の有効性を実証するために日英方向の翻訳でモデルやデコーダの設定を変化させ，その改善後に同定結果がどのように変化するかを検証した．検証結果からは提案手法によって統計的機械翻訳システムの課題をある程度有効に表現できることがわかった．

Automatic Diagnosis Method of Translation Errors in Statistical Machine Translation Systems

Katsuhiko Hayashi† Seiichi Yamamoto†
†Doshisha University

ABSTRACT

In this paper, we propose a method to diagnose causes of translation error in statistical machine translation (SMT) system as language model, translation model, or search problem. This method estimates the cause of translation error by comparing score of the translation with that of a reference sentence. We testify how the diagnosis of translation errors changes, as models or decoder in the SMT system. As a result, it has been confirmed that the proposed method is effective to diagnose translation errors.

1. はじめに

近年，世界のグローバル化に伴い，母語ではない他の言語（L2）でのコミュニケーション機会が増加してきており，コンピュータによって異言語間の翻訳を行う機械翻訳技術が重要なものとなってきている．機械翻訳技術には様々な手法が存在するが，近年のコーパスの開発の増加に伴い，統計的機械翻訳（Statistical Machine Translation）が活発に研究されており，一定の成果を上げてきている[1]．しかし，現実世界への広範な適用を考える際には，その精度はまだ十分に満足できるものではなく改善の余地は大きい．

SMTは対訳コーパスや目的言語コーパスを用いて翻訳モデルと言語モデルを学習しておき，入力文に対してそれらの確率モデルの出力値が大きくなる目的言語文を探索するというシステムである．確率モデルにおける問題は，データスパースネス問題などにより，入力文に対して本来望ましい翻訳解に正しい値を割り当てることができないといったことが挙げられる．一方，モデルに基づいて解を探索するデコーダはNP-Completeであることが知られており，近似的な探索手法をとることでデコーディングが行われる．そのため，探索過程で行われる枝刈りなどの作業により，本来正しいとされる解が探索できないといった問題が生じることになる．

特定の統計翻訳システムにおいて確率モデル，あるいはデコーダの問題がどの程度翻訳誤りを引き起こす原因となっているかを知る方法は現在まで考案されておらず，従来は直感的にシステムの問題改善がなされてきた．しかし，本来はシステムのどの部分にどの程度問題があるかを知った上で改善を行う方がはるかに効率良く問題解決が行える．そこで本稿では統計翻訳システムにおいて翻訳誤りの原因となる箇所として言語モデル，翻訳モデル，デコーダの3つを考え，個々の翻訳事例に対して

これらのうちでどの部分が翻訳誤りの原因となったかを自動同定する手法を提案する．従来，翻訳機のシステム性能はBLEU[2]などに代表されるように翻訳結果と参照訳（リファレンス）との類似度に基づいて翻訳自動評価した結果から比較検証されてきた．しかし，本稿の提案手法では従来の評価手法[3]とは異なり，統計翻訳機の問題を直接的に自動同定し，定量化できるという利点がある．

2. 統計的機械翻訳

統計翻訳ではある原言語の入力文Jを目的言語のテキストEに翻訳する問題を最大尤度の解を発見する問題としてとらえる．

$$\hat{E} = \arg\max_{E} P(E \mid J)$$

さらにベイズの定理から上式は最大事後確率を求める問題に置き換えられる．

$$\hat{E} = \arg\max_{E} P(E)P(J \mid E)$$

ここで$P(E)$を言語モデル，$P(J \mid E)$を翻訳モデルと呼ぶ．また，これらのモデル尤度に基づいて解を探索する過程をデコーダと呼ぶ．

言語モデルは目的言語のテキストが文章としてどれだけ確からしいかという訳語並びを評価するモデルである．言語モデルには一般的にN-gramモデルが使われ，目的言語のコーパスから学習される．一方，翻訳モデルは目的言語と原言語のテキストがあったときに，その対がどれだけ翻訳として確からしいかを評価するためのモデルである．翻訳モデルの学習には原言語と目的言語のテキストが対となった対訳コーパスが利用され，単語単位の翻訳モデルではIBMモデル1～5[4]が一般的である．デコーダでは基本的に訳語を最適な順序に並び替える問題を解くことになる．これは巡回セールスマン問題と等価であり，NP-Completeな問題である．統計翻訳におけるデコーダの探索手法は様々に存在するが代表的なものとしては初期近似解から仮説を探索するグリーディアルゴリズム[5]や入力文に対して出力文を文頭から生成していくようなビームサーチアルゴリズム[6][7]などがある．

3. 翻訳誤り原因自動同定手法

統計翻訳の主要構成要素は言語モデル，翻訳モデル，デコーダの3つである．そこで，ある翻訳事例が翻訳誤りを引き起こす原因もそれら3つが主なものであると考えるのが妥当である．

本稿では誤り原因を自動同定する方法として図1のような決定木による手法を提案する．提案手法の基本的な概念は音声認識システムへの適用例として提案されたもの[8]と同一であるが，音声認識とは異なり機械翻訳では解が一意に決定されないため，バイリンガルによって作成された複数の参照文を正解とする．

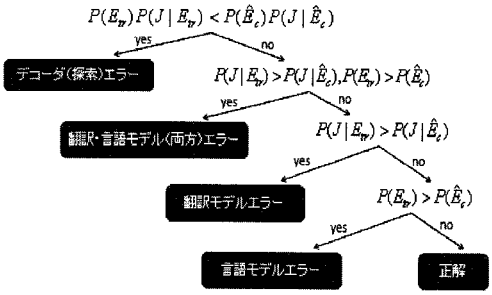

図1 翻訳誤り原因同定のための決定木

ここでJとは原言語入力文のことであり，ある統計翻訳機によるその翻訳結果をE_{tr}とする．また，Jに対する参照訳をS種類用意し，$E_c^s(1 \le s \le S)$と定義する．さらに，S種類の参照訳の中から翻訳モデルと言語モデルの積が最大となる参照訳を$\hat{E}_c$とする．これより翻訳誤り原因を自動同定する手続きは図1の決定木に従う．

図1の決定木における最初の分岐では翻訳結果の言語・翻訳モデルの積と最大尤度を持つ参照訳の言語・翻訳モデルの積を比較し，参照訳の尤度の方が高い場合は探索の問題であるとしている．これはモデルが理想的な解とされる参照訳に高い値を割り当てているにも関わらず，その解を探索することができなかったデコーダの問題であるとできるからである．一方，最初の分岐で翻訳結果の尤度の方が参照訳の尤度より高い場合は次以降の分岐でモデルの問題として考える．モデルの問題としては翻訳・言語モデル両方，翻訳モデルのみ，言語モデルのみの3つが考えられ，それぞれ参照訳よりも翻訳結果の方が高い尤度を持つ場合に同定される．これは本来参照訳には高い値が割り当てられるべきであるにも関わらず，翻訳結果の方が高い値をとってしまっているモデルの問題であるとできるからである．また，それら全ての条件にも当てはまらない場合，すなわち，翻訳モデルと言語モデルの値が参照訳と翻訳結果で完全に一致する場合，その翻訳結果は正解であるとしている．

4. 実験と考察

今回，実験に用いた統計的機械翻訳システムは図2に示すようなATRで開発されたグリーディデコーダ方式の統計的機械翻

訳システム[9]である．また，比較検証用にそのモジュールを利用して作成したビームサーチデコーダも使用する．デコーディングアルゴリズムは渡辺らによる文献[10]を参考とした．言語モデルはN-gram，翻訳モデルはIBM モデルであり，実装は3-gram，IBM モデル4である．

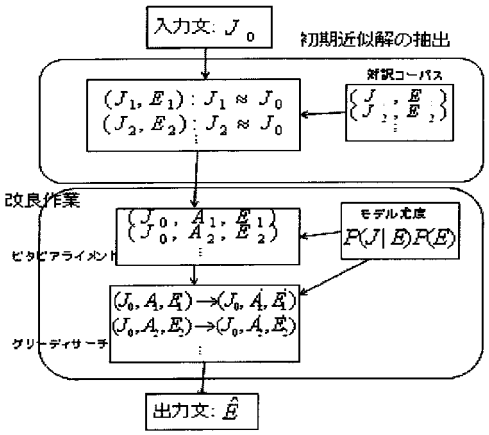

図2 実験に使用するグリーディデコーダ（SAT）

次に，実験に用いるテストセットはATRで開発された日英旅行対話コーパスBTEC（Basic Travel Expression Corpus）[11]からランダムに選んだ日本語文90文である．また，日英バイリンガルによって作成された英語参照訳を日本語文に対し各10種類用意した（合計900文）[12]．表1にその1例を示す．また，入力日本語文90文の平均単語長は10.76である．

表1: テストセットの例

日本語文	ひどい雨で彼女は外出することができませんでした。
参照訳①	She could not go out because it was raining hard.
参照訳②	As it was pouring, she could not go outside.
参照訳③	As it was pouring rain, she could not go out.
参照訳④	It was pouring so she was not able to go outside.
参照訳⑤	She couldn't go out because it was pissing rain out.
参照訳⑥	The showers kept her indoors.
参照訳⑦	She stayed in because it was coming down like cats and dogs.
参照訳⑧	The wet weather kept her from going out.
参照訳⑨	The downpour kept her in.
参照訳⑩	Because of the heavy precipitation she stayed inside.

以下では，これらのテストセットを用いて日英方向の翻訳設定とし，提案手法からの実験結果を提示する．

4.1 グリーディデコーダによる実験結果

まず，グリーディデコーダ（SAT）による実験結果を提示する．表2では翻訳機の実験設定を示しており，実験手順はそれに従うこととする．まず，実験設定①による実験結果を図3に示す．

表2: 翻訳機の実験設定

	言語モデル	翻訳モデル	デコーダ
実験設定①	BTEC 英語コーパス 16 万文（3-gram）	BTEC 日英対訳コーパス 16 万文から学習	改良手続きが収束するまで
実験設定②	英語参照訳900文（3-gram）と BTEC 英語コーパス 16 万文（3-gram）の線形補間モデル	BTEC 日英対訳コーパス 16 万文から学習	改良手続きが収束するまで
実験設定③	英語参照訳900文（3-gram）と BTEC 英語コーパス 16 万文（3-gram）の線形補間モデル	BTEC 日英対訳コーパス 16 万文から学習	改良手続きを 5 回に制限

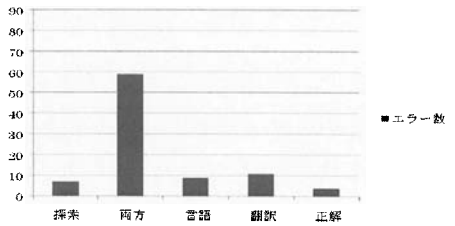

図3 設定①による実験結果

図3からは設定①における翻訳においてモデルの問題が非常に多くの割合を占めていることがわかる．これは学習に用いたコーパスが言語モデル，翻訳モデル共に16万文と少なかったことが理由に挙げられる．すなわち，モデルの精度がよくないため，参照訳に高い尤度が割り当てられなかったからだと言える．また，正解となる例が4文存在するが，実験に用いているグリ

ーディデコーダは学習データとして使用している BTEC16 万文中に入力文と Exact Match する原言語文があれば，その対訳を答えとしているため，テストセットに学習データとして使われたものがある場合は同定結果として翻訳結果が正解となる可能性は高い．

次に実験設定②では言語モデルの問題にのみ着目し，その改善を行った場合の結果を示す．改善方法としては入力文に適したコーパスを収集してモデルを学習することが 1 つの方法であると考えられ，ここでは入力文に対する参照訳計 900 文を用いて言語モデルを学習し，BTEC16 万文から学習したモデルとの線形補間モデルを作成した．結合重みは参照訳計 900 文から作成したモデルに 0.8，もう一方に 0.2 で結合を行った．設定②の結果を図 4 に示す．

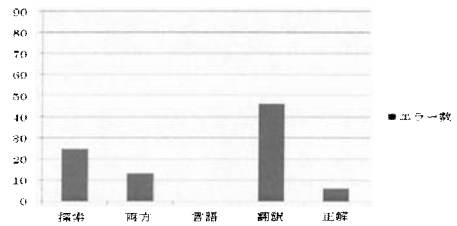

図 4 設定②による実験結果

図 4 からは言語モデルの問題が大きく減少していることがわかる．これは参照訳に対して最適化した言語モデルであるため，参照訳の言語モデル尤度が翻訳結果の言語モデル尤度を上回ったからであると考えられる．また，入力文に対して理想的な言語モデルを実装したことにより，図 3 と比較して正解の数が多少増加していることがわかる．

次に実験設定③では設定②とは異なり，デコーダの探索空間を縮小した．すなわち，デコーダの性能を下げることで結果にどのような変化が出るかを調べる．結果を図 5 に示す．

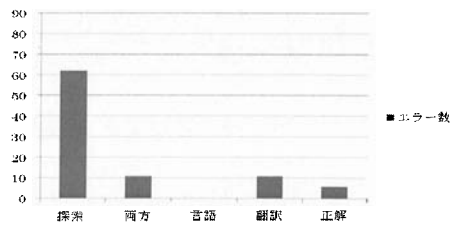

図 5 設定③による実験結果

図 5 の結果からは図 4 と比較し，翻訳機の問題が探索の問題の方へ流れていることがわかる．これは翻訳機のデコーダの問題が大きく出たからであると考えられるが，本質的に翻訳モデルの問題は残ったままである．これは提案手法が統計翻訳機のより大きな問題を持つ部分に原因が引っ張られているからであると考えられる．

4.2 ビームサーチデコーダによる実験結果

ここではビームサーチデコーダによる実験結果を提示する．ビームサーチデコーダはグリーディデコーダと異なり，初期近似解を必要としないので，そのバイアスの影響を受けない．そのため，デコーダの手続きをよりはっきりさせて検証することができると考えられる．表 3 では翻訳機の実験設定を示しており，実験手順はそれに従う．実験設定①による実験結果を図 6 に示す．

表 3: 翻訳機の実験設定

	言語モデル	翻訳モデル	デコーダ
実験設定①	BTEC 英語コーパス 16 万文（3-gram）	BTEC 日英対訳コーパス 16 万文	ビーム幅 20
実験設定②	英語参照訳 900 文（3-gram）と BTEC 英語コーパス 16 万文（3-gram）の線形補間モデル	BTEC 日英対訳コーパス 16 万文	ビーム幅 20
実験設定③	英語参照訳 900 文（3-gram）と BTEC 英語コーパス 16 万文（3-gram）の線形補間モデル	BTEC 日英対訳コーパス 16 万文	ビーム幅 100

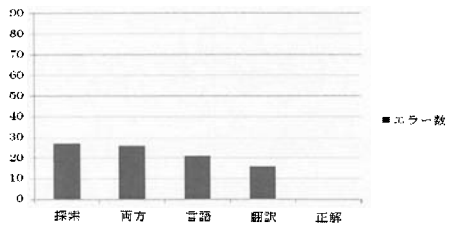

図 6 設定①による実験結果

図6からは全ての問題がある程度均等に起こっていることがわかる．これは図3のグリーディデコーダでの検証時とは異なり，デコーダに多くの問題があるからであると考えられる．

実験設定②ではグリーディデコーダのときと同様に言語モデルの改善を行い，実験を行った．その結果を図7に示す．

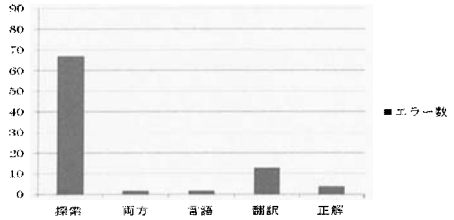

図7 設定②による実験結果

この結果からは言語モデルの問題が大きく減少していることがわかる．また，翻訳モデルとデコーダを比較して，より大きな問題である探索の方に原因同定が多くなされていることがわかる．図6と比較し，言語モデルの改善後には正解が増加していることもわかる．

次に図6,7からデコーダの問題が大きいと考えられるので，仮説を保存するキューの幅を100に増加させた．これはより幅広い空間を探索できるようにしたことを意味し，デコーダの性能が向上したと考えてよい．これによる実験結果を図8に示す．

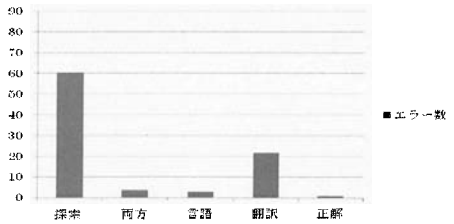

図8 設定③による実験結果

図8からはデコーダの探索範囲を広げたことにより，探索の問題が減少していることがわかる．これはデコーダの性能が上がったことにより，より良いモデル尤度を持つ解が探索できるようになったからであると考えられる．

4.3 実験結果に対する考察

実験では言語モデルやデコーダの変化によって，結果がどのように変化するかを検証した．その結果，言語モデルがよくなれば言語モデルの問題は減少し，デコーダの性能によって探索の問題が適切に増減することがわかった．これより提案手法は統計的機械翻訳システムの状況をある程度説明できていると考えられる．しかし，一方で，言語モデルを大きく改善しても正解となる数はそれほど増加していないことがわかる．これはデコーダや翻訳モデルにまだ問題が残っていることと同時に，提案手法では複数の参照訳の中で最大の尤度を持つ参照訳と翻訳結果が一致したとき以外には正解にならないという条件があるからである．ここで，図3の結果において正解となっているものの1例を表4に示す．

表4：正解となる例

入力文：転んだ。
翻訳結果：I fell down.
参照訳：I fell down.

表4で示すように正解となる例の多くはテキストとして短文なものが多い．これは長い文章になると参照訳と翻訳結果が全く同じになるということがほとんどあり得ないためである．これは他の実験設定における正解についても同じことが言える．

実験時の手法では比較に用いる参照訳を選び出す際に最大尤度のものを選び出していたが，残りの9文については全く考慮していなかった．しかし，残り9文の中に翻訳結果と一致するものが存在する可能性があるので，残りの参照訳について翻訳結果と完全に一致するものがあるかどうかを調べた．その結果を表5に示す．

表5：参照文の選択による正解数の違い

	実験設定①	実験設定②	実験設定③
グリーディ（最大尤度参照文のみ）	4/90 文	6/90 文	6/90 文
グリーディ（参照文10文全て）	4/90 文	9/90 文	9/90 文
ビームサーチ（最大尤度参照文のみ）	0/90 文	4/90 文	1/90 文
ビームサーチ（参照文10文全て）	1/90 文	5/90 文	1/90 文

表5の結果からは最大尤度の参照文とではなく残りの参照文と完全に一致する翻訳結果になっているものがいくつか見られるが，やはり，正解の数としては少ないと言える．また，正解の

多くはやはり 3 単語以下の短い文章であり，長い文章が完全に一致することはかなり多くの参照訳を集めても難しいと考えられる．

そこで翻訳自動評価指標のひとつである Word Error Rate (WER) [13]を用いて各翻訳機設定における訳質評価を表 6 に提示する．WER とは DP (Dynamic Programming) により，置換，挿入，削除のペナルティの総和を最小化するアライメントを算出し，翻訳結果文と参照訳を単語単位で比較して評価するものである．よって，この尺度からは翻訳結果が参照訳にどれほど文字列として近づいているかをみることができる．そのため，正解が増加していなくても，翻訳結果が参照訳に近づいているかを WER の値から知ることができると考えられる．また，ここでは複数の参照訳を用いた m-WER (multiple reference WER) で計算する．ちなみ，WER は値が低いほど単語誤りの割合が少なく，良い値ということになる．

表 6: m-WER による翻訳自動評価の結果

	実験設定①	実験設定②	実験設定③
グリーディ	51.1	44.2	44.3
ビームサーチ	54.7	49.0	49.7

表 6 における実験設定①と②の比較から，言語モデル改善後には両デコーダとも翻訳結果が参照訳に近づいていることがわかる．しかし，提案手法では正解の数が増加しない限り，翻訳システムには必ず何らかの問題があるとされるため，翻訳システムのこのような性能向上を正確に表現することができていないという問題が指摘できる．

最後にグリーディデコーダとビームサーチデコーダ間での考察を行う．表 2，表 3 の設定①による実験結果図 2，図 5 を比較するとビームサーチデコーダの方が探索の問題が多いことがわかる．これは実験に使用したビームサーチデコーダの手続きを簡易的に作成したことビーム幅を非常に狭くとっていたことが原因であると考えられる．提案手法の結果からは検証に使用したグリーディデコーダではモデルの改善が優先され，ビームサーチデコーダではデコーダの改善が優先されるということが言える．

5. まとめ

本稿では統計的機械翻訳における翻訳誤り原因を自動同定する手法を提案し，その有効性を検証した．実験では言語モデルとデコーダの状態を変化させることで原因同定結果がどのように変化するかを示した．検証結果では言語モデルの改善後には言語モデルの問題が減少し，デコーダの探索空間を広げた際には探索の問題が減少し，探索空間を縮小した際には探索の問題が増加するという傾向が得られた．これは提案手法が統計翻訳機の状態をある程度正確に評価していると考えてよい．

実験では言語モデルの改善手法として参照訳 900 文を用いて作成した言語モデルを線形補間するということを行ったが，今後はより一般性のあるやり方で言語モデルの改善を行う必要があると考えられる．ドメインに適した言語モデルを線形補間するという観点で言えば，言語モデルを単語の素性からクラスタ化しておき，入力文に適したクラスの言語モデルを逐次的に線形補間するといったことが考えられる．一方，翻訳モデルやデコーディング手続きの改善は今回行わなかったが，現在，盛んに研究が進められているフレーズベースの統計的機械翻訳システム[14][15]を用いて同様の検証を行うことは重要な課題であると考えられる．

最後に，提案手法ではどれだけ高精度な統計的機械翻訳システムを用いて検証したとしても正解が大きく増えることはないと考えられる．つまり，提案手法は特定の統計的機械翻訳システムの問題点を定量的に表現することはできるが，その性能を評価することはできないという問題がある．また，問題点を定量的に表すことにおいても，言語モデルが文長によって尤度が変化してしまうという問題により，高精度な尤度比較は行えていないのが現状である．音声認識システムにおける適用例では文章全体の尤度比較ではなく誤り区間を特定し，部分による尤度比較を行っていたが，正解を一意に限定することのできない機械翻訳ではそのような比較は困難であると言える．今後，統計的機械翻訳においても部分区間での比較を行うことを検討する場合には参照訳の選択方法として翻訳結果と参照訳の編集距離を計算し，文字列として最も近い参照訳を選び出す方法が考えられ，検証を行う必要があると言える

謝辞

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．本研究に際してグリーディデコーダ方式及びビームサーチデコーダ方式について御教示頂いた渡辺太郎氏 (NTT CS 研究所) 及び安田圭志 (ATR) に感謝します．

参考文献

[1] Paul, M.: Overview of the IWSLT 2006 evaluation campaign, *Proc. International Workshop on Spoken Language Translation (IWSLT)*, pp. 1-15 (2006).

[2] Papinei, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proc. ACL 2002*, pp.311-318 (2002).

[3] 安田 圭志, 隅田 英一郎: 機械翻訳の研究・開発における翻訳自動評価技術とその応用, 人工知能学会誌, Vol.23, No.1, Jan.2008

[4] Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, 18, 4, pp. 263–311 (1993).

[5] Germann, U., Jahr, M., Knight, K., Marcu, D., Yamada, K.: Fast Decoding and Optimal Decoding for Machine Traslation, *Proc. ACL 2001*, Toulouse, France (2001).

[6] Och, F.J., Ueffing, N., Ney, H.: An Efficient A* Search Algorithm for Statistical Machine Translation, *Proc. ACL-2001 Workshop on Data-Driven Machine Translation*, Toulouse, France, pp.55-62 (2001).

[7] Wang, Y.-Y., Waibel, A.: Decoding Algorithm in Statistical Machine Translation, *Proc. ACL 1997* (1997).

[8] 南條 浩輝, 李 晃伸, 河原 達也: 大語彙連続音声認識における認識誤り原因の自動同定, 音声言語情報処理, Vol.99, No.64, SLP-27-6 , Jul.1999.

[9] Watanabe, T., Sumita, E.: Example-based Decoding for Statistical Machine Translation, (2003)

[10] 渡辺 太郎, 隅田 英一郎, 奥乃 博: 生成方向を考慮した統計的機械翻訳のためのデコーディングアルゴリズム, 情報処理学会論文誌 , Vol.44 , No.12 , Dec. 2003.

[11] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., Yamamoto, S.: Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, *Proc. LREC 2002*, Las Palmas, Canary Islands, Spain, pp.147-152 (2002).

[12] 喜多村 圭祐, 安田 圭志, 山本 誠一, 柳田 益造: 日本人英語学習者による日英翻訳コーパスの開発, 電子情報通信学会, 2007.

[13] Leusch, G., Ueffing, N., Ney, H.: A novel string-to-string distance measure with applications to machine translation evaluation, *Proc. MT Summit IX* (2003).

[14] Richard Zens, Franz josef Och, and Hermann Ney. Phrase-Based Statistical Machine Translation., *KI*, 2002.

[15] Taro Watanabe, Eiichiro Sumita, and Hiroshi G. Okuno. Chunk based Statistical Translation, *41st Annual Meeting of the Association for Computational Linguistics*, 2003.