Internal Use Only（非公開）

TR－SLT－0092

複数の編集距離を用いた口語翻訳文の自動評価

Automatic Grader of MT Outputs in Colloquial Style by Using Multiple Edit Distances

秋葉　泰弘
Yasuhiro AKIBA

2005年3月25日

概要

本稿では，翻訳文の品質を評価するという人間の知的能力を自動化する挑戦的で新しい試みを報告する．ここで翻訳文は，機械翻訳(MT)システム，特に会話を翻訳する音声翻訳(SSMT)システムの翻訳部からの出力である．従来法にはBLEUがあるが，SSMTシステムの評価には，BLEUは以下の2つの理由で不向きである．第1に，誤りの評価はその出現箇所に依存するべきでないが，BLEUは，文頭で起きた誤りは軽く評価し，文中では重く評価する．第2に，BLEUは話し言葉を処理する上で寛容さに欠ける．BLEUは，誤りに阻害されずに会話を続けることができるような些細な誤りも許さない．著者は，複数の編集距離を用いて翻訳文に自動的に評点を付与する新自動評価法REDを報告する．REDでは，訓練事例から翻訳文に自動的に評点を付与する決定木を学習し，この決定木を用いて評点が未知の翻訳文に対して評点を付与する．各訓練事例は翻訳文と人手評点の対の集合であり，翻訳文は複数の編集距離を用いてベクトル化される．ここで複数の編集距離は，通常の編集距離(ED)およびEDの拡張版である．新たな評価対象の翻訳文は，訓練事例と同様にベクトル化を行ない，学習した決定木を用いて評点を付与する．REDの評価は誤りの出現箇所に依存しない．また，これら複数の編集距離を用いることで，REDの評価はED単体やBLEUよりも些細な誤りに対して寛容となる．REDとBLEUの性能比較のために，これらでMTシステムを評価する実験を行ったところ，REDはBLEUより性能が良いことが示された．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

1 はじめに

本稿では，翻訳文の品質を評価するという人間の知的能力を自動化する挑戦的で新しい試みを報告する．ここで翻訳文は，機械翻訳 (**MT**) システム，特に会話を翻訳する音声翻訳 (**SSMT**) システムの翻訳部，からの出力とする．近年，対訳コーパス (原文とその対訳の対からなるコーパス) から翻訳知識を自動的に学習する技術が進展し，それに伴いMT システムの研究者や技術者の間で MT システム評価の自動化を望む声が高まりつつある．MT システムの自動評価が実現されれば，新しい MT システム開発手法の有効性を敏速に検証できる．本稿ではSSMT システムの評価基準として，ATR(国際電気通信基礎技術研究所) が提案した翻訳文評価基準 [Sumita 99] を用いる．人間の評点者が翻訳文を評価する場合には，同評価基準で定義されている 4 段階の評点[1] A〜D の中から 1 つを選び，翻訳文に付与する．この付与の自動化が本稿で報告する試みである．

翻訳評価の自動化への期待に応じて，近年，自動翻訳評価法が提案され始めた．その中に BLEU[Papineni 01] がある．BLEU は N グラム正解率に基づいて翻訳の良さを測る．即ち，翻訳文中に含まれる連接する N 個の単語 (N グラム) を評価の単位とし，翻訳文中のすべての N グラムのうち何個が**参照訳** (人間により正しく翻訳された訳文) に含まれているかに基づいて，翻訳の良さを測る．BLEU は多くの MT システム研究者の注目を浴び，新しい MT システム開発手法の良さを検証するために BLEU が使用されている [Yamada 02, Och 02, Marcu 02].

本稿で対象とするSSMT システムの評価には，BLEU は以下の 2 つの理由から不向きである．第 1 に，2 章で論じるように，誤りの評価はその出現箇所に依存するべきではないにもかかわらず，BLEU では評価値が誤りの出現箇所に依存する．具体的には，文頭や文末で起きた誤りは軽く評価され，文中では重く評価される．誤り評価の出現箇所依存性は，翻訳文が短い場合に特に顕著となる．話し言葉の平均文長は，話し言葉コーパスの多くに見られるように短い [国研 55]．例えば，ATR conference registration corpus や DARPA resource management corpus[Kitano 94] では 7〜9 単語，Verbmobile corpus[Zens 03] では 10 単語前後，ATR BTEC corpus[Takezawa 02] では 6〜8 単語である．

第 2 に，BLEU は話し言葉の文を処理する上で寛容さに欠ける．話し言葉の文は文法的に誤りが含まれることがあるが，我々はそのような発話を理解し，誤りに阻害されずに会話を続けることができる．一方 BLEU は，いかなる誤りも許さず，些細な誤りにも評価上ペナルティーを課す．些細な誤りとしては例えば，(1) 模範的な訳語の代わりに類義語を選択し翻訳文が流暢でない，(2) 修飾語の並び順が模範的でなく翻訳文が流暢でない，(3) 重要でない単語が翻訳文中に現れていない，等である．SSMT システムを評価する評価法は，このような些細な誤りに対して寛容である必要がある．即ち，上記のような些細な誤りを許容し，我々が会話を容易に続けることができないような致命的な誤りにのみ評価上ペナルティーを課すべきである．

[1] ATR の翻訳文評価基準 [Sumita 99] は以下の 4 段階で定義される：(A) 優: すべての情報は訳出され，文法的にも問題無い．(B) 良: 重要でない情報が多少訳出されていないまたは文法的にも多少問題があるが，原文の情報を容易に復元できる．(C) 可: 重要でない情報が多数訳出されていない，または文法的にも大分問題があるが，努力をすれば原文の情報がなんとか復元できる．(D) 不可: 重要な情報が訳出されていないか誤訳されており，原文の情報を復元できない．

本稿では，複数の編集距離を用いて翻訳文に自動的に評点(前記4段階評点[1]の1つ)を付与する新自動評価法，RED (g**R**ader based on **E**dit **D**istances)[秋葉 05, Akiba 01, Akiba 03]を報告する．具体的には新自動評価法では，翻訳文に自動的に評点を付与する決定木を訓練事例から学習し，この決定木を用いて評点が未知の翻訳文に対して評点を付与する．各訓練事例は翻訳文と人手評点の対であり，翻訳文は複数の編集距離を用いてベクトル化される．ここで複数の編集距離は，挿入(insertion)，削除(deletion)，置換(replacement)を編集操作とする通常の編集距離(**ED**)および，EDの拡張版である．拡張版の編集距離は，以下の4つの拡張方針またはそれらの組み合わせに対応する．同拡張方針は，(1) 交換(swap)を編集操作として新たに使うか否か，(2) 内容語について意味コード[2]の参照を許すか否か，(3) 編集単位を内容語に限定するか否か，(4) 編集単位をキーワードに限定するか否か，である．ここで編集単位とは，通常の編集距離同様に挿入や削除等の編集操作を行なう操作対象の候補である．これら複数の編集距離を用いることで，REDはED単体やBLEUよりも些細な誤りに対して寛容となる．決定木は一般に，判断結果を表すクラスラベルに基づいて学習される．この学習に利用される各翻訳文のクラスラベルは複数人による人手評点のメジアン[3]とする．新たな評価対象の翻訳文は訓練事例と同様にベクトル化され，学習した決定木を用いて評点を付与される．

REDとBLEUを実験的に比較した．REDでは各翻訳文の訳質を評価することに主眼が置かれているのに対し，BLEUではMTシステム同士の性能を比較することに主眼が置かれている．そこで著者は，それぞれの主眼に則した以下の2種類の評価実験を，さまざまな性能のMTシステム群を用いてATRが構築した会話文の大規模な対訳コーパスBTEC[Takezawa 02]上で行なった．第1の実験では，REDやBLEUで翻訳文に評点を付与する(以下，これを文レベル評価と呼ぶ)際の正解率を比較する．第2の実験では，REDやBLEUで1つのMTシステムが別の翻訳システムより有意に優るか否かを判定させる(以下，これをシステムレベル評価と呼ぶ)際の正解率を比較する．両実験において，REDはBLEUより性能が良いことが示された．

以下，次章ではBLEUの誤り評価値が誤りの出現箇所にいかに依存するかを例を用いて論じる．3章で新自動評価法REDを報告する．実験結果とその考察を4章で示し，関連研究について5章で議論する．最後に6章でまとめる．

2 Nグラムを用いた評価の問題点

本章ではまずBLEUを概説する．次に前章で簡単に述べた，BLEUをはじめとするNグラムを用いた評価の誤り出現箇所依存性について論じる．

BLEUは，ユニグラム，バイグラム，トライグラム，4グラムの各Nグラム正解率[4]を基に自動評価を行なう．Nグラム正解率は，複数の参照訳(1章参照)に対して計算される．

[2]意味コードとしては，単語の意味を指し示すものであればなんでもよい．本稿では[大野 81]記載の意味コードを利用した．

[3]メジアンを採用する理由は，多くの人手評点はメジアンと一致すると期待され，翻訳文の平均的な評価でラベル付けできるからである．

[4]厳密な定義は，原著論文[Papineni 01]を参照していただきたい．

参照訳 $\hat{R}_1^9$:	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
誤訳 T_1:	R_1	$\boxed{R_2'}$	R_3	R_4	R_5	R_6	R_7	R_8	R_9
誤訳 T_2:	R_1	$\boxed{R_0}$	R_2	R_3	R_4	R_5	R_6	R_7	R_8 R_9
誤訳 T_3:	R_1	$\square$	R_3	R_4	R_5	R_6	R_7	R_8	R_9

図 1: 参照訳 $\hat{R}_1^9$ の2番目の位置に誤りがある誤訳の例 T_1，T_2，T_3：T_1 では，2番目の語 R_2 が他の語 R_2' に誤って置換されている．T_2 では，2番目の位置に語 R_0 が誤って挿入されている．T_3 では，2番目の語 R_2 が削除されている．

表 1: 誤りの評価がその出現箇所に依存することを示す3つのパターン：(パターン 1) 位置 i の語が誤って他の語に置換されている．(パターン 2) 位置 i に語が挿入されている．(パターン 3) 位置 i の語が削除されている．

出現箇所 i	1	2	3	4	5	6	7	8	9
(パターン 1)	1/6	2/6	3/6	4/6	4/6	4/6	3/6	2/6	1/6
(パターン 2)	1/7	2/7	3/7	4/7	4/7	4/7	4/7	3/7	2/7
(パターン 3)	0/5	1/5	2/5	3/5	4/5	3/5	2/5	1/5	0/5

例えば4グラム正解率は，翻訳文に含まれる4グラムのうち，いずれかの参照訳に含まれる4グラムの割合である．ここで“いずれかの参照訳”とあるが，これは，すべての4グラムを同一の参照訳と照合する必要はなく，各4グラムが別々の参照訳と照合してもよいことを意味する．

Nグラムを用いた評価法が誤りをどう評価するかを図1で説明する．誤訳 T_1，T_2，T_3 のおのおのには，参照訳 $\hat{R}_1^9$ の2番目の位置に誤りがある．誤訳 T_1 に含まれる2つの4グラム “$R_1\ R_2'\ R_3\ R_4$” と “$R_2'\ R_3\ R_4\ R_5$” は，参照訳 $\hat{R}_1^9$ にない．同様に，誤訳 T_2 に含まれる2つの4グラム “$R_1\ R_0\ R_2\ R_3$” と “$R_0\ R_2\ R_3\ R_4$”，および誤訳 T_3 に含まれる1つの4グラム “$R_1\ R_3\ R_4\ R_5$” は，参照訳 $\hat{R}_1^9$ にない．従って，他のどの参照訳にもこれら5つの4グラムが含まれていなければ，各4グラム誤りの評価値 (= 1 − ‘4グラム正解率’) は，各誤訳について順に 2/6，2/7，および 1/5 となる．

表1は，4グラム誤りの評価値が誤り出現箇所に応じてどう変化するかを示す．4グラム誤りの評価値では，誤りの出現箇所を含む4グラムの数だけて誤りを重複して数え上げるため，文中で誤りを数え上げ回数は4回であるのに対し，文頭や文末近くではより少ない回数しか数え上げられない．それゆえに，文頭や文末に近いほど4グラム誤りの評価値は小さい．この現象はバイグラムやトライグラムでもまた同様に現れる．誤り評価の出現箇所依存性は一般に望ましくない．誤り評価の不均一さは，対話翻訳のように，翻訳文が短い場合[5]に特に顕著となる．

[5]文書翻訳におけるように翻訳文が長い場合には，文中に現れる殆んどの誤りは，例えば表1の (パターン 1) における 4/6 のように，等しい重みで評価される．

3 自動評価法：RED

本章では，複数の編集距離を用いて翻訳文に自動的に評点を付与する新自動評価法，RED (g**R**ader based on **E**dit **D**istances)[秋葉 05, Akiba 01, Akiba 03] を報告する．REDは以下の2種類の情報を利用する．1つは少なくとも3名以上の評点者により評価されたMTシステムによる翻訳文であり，もう1つは複数の参照訳 (1章参照) である．前者の情報を利用する理由は，REDが翻訳文に付与する評点としては複数人が評価した場合の平均的な評点が望ましいからである．一方，後者の情報を利用する理由は，[Thompson 91] や [King 96] で述べられているように，入力文に対する翻訳文は1通りとは限らず，数多くの翻訳文が等しく容認され得る点にある．

3.1 REDの概要

REDは学習フェーズと評価フェーズからなる．学習フェーズでは，評点を付与する決定木[6]を学習し，評価フェーズでは，学習された決定木を用いて評価対象の翻訳文に評点を付与する．

学習フェーズは，以下の3つのステップ (L1)，(L2)，および (L3) からなる．

(L1) 決定木の訓練事例を作成するために，MTシステムによる翻訳文に**メジアン評点**を付与する．ここでメジアン評点とは，予め複数人間評点者[7] がおのおの付与した評点のメジアンである．例えば，MTシステムによる翻訳文を2名の評点者が "A"[1] と評点し，別の1名が "C" と評点した場合には，この翻訳文は評点 "A" が付与される．また，1名の評点者が "A" と評点し，別の1名が "B" と評点し，更に別の1名が "C" と評点した場合には，翻訳文は評点 "B" が付与される．

(L2) 上記の (L1) でメジアン評定を付与された各翻訳文を，17次元のベクトルとして表す．第1成分は通常の編集距離 (ED) の値であり，末尾の成分はその翻訳文に付与されたメジアン評点である．その他の成分は，EDを拡張した15種類の編集距離 (詳細は次節で述べる) のいずれかの値である．例えば，図2の入力文 (S)，MTシステムによる翻訳文 (T)，3つの参照訳[8](R1)〜(R3) を用いて第1成分の値の求め方を説明すると，翻訳文と各参照文の対，{(T), (R1)}, {(T), (R2)} および {(T), (R3)} に対してEDを計算し，その最小値を第1成分とする．上記のその他の成分の値も，同様に計算する．

(L3) (L1) および (L2) を通して準備した訓練事例を用いて，[キンラン 95] などの決定木学習アルゴリズムにより，決定木を学習する．

---

[6]教師有り学習アルゴリズムの代表例として，本稿では決定木学習アルゴリズムを用いた．決定木学習アルゴリズムは，多くの現実問題への適用実績のある教師有り学習アルゴリズムである [秋葉 98]．

[7]学習フェーズの説明では説明を簡単にするために評点者の数を3名としたが，後述するように評価実験ではメジアン評点の信頼性を確保するために9名とした．

[8]参照訳はできるだけ多く (最低3つ) 事前に準備しておく．参照訳を網羅的に準備することは難しいため，REDにおけるEDの拡張版のあるものでは，参照訳を語彙的に換言したり，構造的に換言したりする．

(S) They are a couple coming to Japan for sightseeing.
(T) 彼らの日本に来るカップルです.
(R1) 彼らは日本へ観光に来るふたりです.
(R2) 観光しに日本に来るカップルです.
(R3) 観光で日本に訪れるふたりです.

図 2: 原文 (S) に対する，英日 MT システムによる翻訳文 (T) および 3 つの参照訳 (R1〜R3) の例．参照訳は，原文 (S) に対する英語の等価表現.

評価フェーズは，次のステップ (E1) および (E2) からなる.

(E1) 上記のステップ (L2) と同様に，MT システムの各翻訳文を 17 次元のベクトルとして表す．ただし，末尾の要素は評点が判らないので NULL とする.

(E2) 学習フェーズで学習した決定木を用いて評価対象の各翻訳文に評点[1](A〜D のいずれか) を付与する.

3.2 編集距離

RED では，前節で説明したように，訓練事例の翻訳文や評価対象の翻訳文を 17 次元のベクトル (前節，L2 参照) として表現する．本節では，その際に利用する編集距離 ED を拡張した 15 種類の編集距離について述べる.

編集距離 ED のすべての拡張版は，編集距離 ED[Wagner 74, Thompson 91] と同様に，挿入 (insertion)，削除 (deletion)，置換 (replacement) を編集操作として利用する．編集操作 (後述の交換も含む) には編集コストを自由に設定できるが，本稿では編集コストをすべて等しく 1 とする．編集距離 ED およびその拡張版の値は翻訳文中の誤りの出現箇所に寄らず誤りが編集コスト 1 で修正されるため，RED の評価は誤りの出現箇所に依存しない.

これらの編集距離 ED および拡張版の編集単位は形態素である．2 つの形態素がマッチする条件は，両形態素の標準形および品詞が共に一致する場合に限る．例えば，表 2 が示すように，(T) と (R1) の対に対して編集距離 ED を計算する場合には，(T) 中の “観光” と (R1) 中の “観光” は標準形と品詞が共に同じなのでマッチするが，(T) 中の “の” と (R1) 中の “観光” は標準形と品詞共に異なるのでマッチしない．出現形の代りに標準形を参照することにより，活用誤り等の些細な誤りが許容される.

次に ED の拡張版のうち，基本となる 4 つの拡張版，ED_{swp}, ED_{sem}, ED_{cnt}, および ED_{key} を順に説明する.

ED_{swp} では，交換操作 (swap)[Lowrance 75, Su 92] を編集操作として新たに使えるように編集距離 ED を拡張する．交換操作を採用すると，“小さな白い花” と “白い小さな花” のように近接した修飾語の順番が入れ替わった場合でも，形態素列が編集コストゼロでマッチできるようになり，マッチングが柔軟となる[9].

[9] この近接の交換を許すと，内容語と助詞の交換が許されるという弊害が懸念されるが，そのような誤り

表 2: 図 2 の翻訳文 (T) および参照訳 (R1)〜(R3) の形態素列の例．* は，その形態素がキーワードと見なされることを示す．

ID	出現形	標準形	品詞	意味コード
T	観光	観光	名詞	892
	の	の	助詞	-
	日本	日本	名詞	709, 719
	に	に	助詞	-
	来る	来る	動詞	**283**, 312
	カップル	カップル	名詞	530
	です	です	BE 動詞	-
R1	彼ら	彼ら	代名詞	892
	は	は	助詞	-
	日本*	日本	名詞	709, 719
	へ	へ	助詞	-
	観光*	観光	名詞	892
	に*	に	助詞	-
	来る*	来る	動詞	283, 312
	ふたり*	ふたり	名詞	530
	です*	です	BE 動詞	-
R2	観光*	観光	名詞	892
	し	する	補助動詞	-
	に*	に	助詞	-
	日本*	日本	名詞	709, 719
	に*	に	助詞	-
	来る*	来る	動詞	283, 312
	カップル	カップル	名詞	530
	です*	です	BE 動詞	-
R3	観光*	観光	名詞	892
	で	で	助詞	-
	日本*	日本	名詞	709
	に*	に	助詞	-
	訪れる	**訪れる**	動詞	**283**, 786
	ふたり*	ふたり	名詞	530
	です*	です	BE 動詞	-

ED_{sem} では，標準形の代りに意味コード[2] を参照し，同じ意味コードを持つ場合には 2 つの形態素をマッチさせる．例えば，(T) 中の “来る” と (R3) 中の “訪れる” は意味コード “ 238” が共通なので，通常の編集距離 ED ではマッチし得ない “来る” と “訪れる” がこの ED_{sem} ではマッチできる．このように ED_{sem} では，標準形が異なる形態素同士でも同じ意味コードを持つ場合[10] には，編集コストゼロでマッチできるようになる．

これら ED_{swp} および ED_{sem} は，近接語の相互交換や同義語置換による語彙的換言により，参照訳 (以下，疑似参照訳と呼ぶ) を構成し，ED で編集距離を計算したのと等価である．疑似参照訳は目的言語の文として流暢さに欠ける可能性があるが，文法的な誤りや情報の欠落がないため A 評点[1] となる．

疑似参照訳に一致する翻訳文は ED_{swp} や ED_{sem} で参照訳との編集距離を計算すると，編集距離がゼロとなる．一方，疑似参照訳と一致する翻訳文の ED の値はゼロより大きく，誤りを含む B 評点以下の訳文も ED の値がゼロより大きいため，これらは ED の値で区別できない．これら拡張版 ED_{swp} および ED_{sem} の意図は，前記疑似参照訳と一致する A 評点の翻訳文と B 評点以下の翻訳文を区別する点にある．

ED_{cnt} では，編集単位をすべての形態素ではなく，名詞や動詞などの内容語に限定する．例えば，(T) 中の “観光” は編集単位であるが，(T) 中の “の” は編集単位ではない．編集単位を内容語に限定することにより，内容語は正しい語順で並んでいるが機能語が若干抜けている場合，例えば，“I went to a zoo” に対して “I went zoo” のような場合，翻訳文は参照訳に編集距離ゼロでマッチする．若干誤りのある翻訳文は，ATR の翻訳文評価基準では評点 B と判断される．

ED_{key} では，編集単位をすべての形態素ではなく，キーワードに限定する．本稿ではキーワードを 2 つ以上の参照訳に現れた形態素と定義する．通常キーワードは内容語のみに限定されるが，この定義では機能語もキーワードになり得る．例えば，図 2 の参照訳 (R1)，(R2)，および (R3) では，“彼ら” は参照訳 (R1) にのみ出現するので “彼ら” はキーワードではない．このキーワードは，ATR の翻訳文評価基準における重要語を想定している．

これら拡張版 ED_{cont} および ED_{key} の意図は，評点が B か C の翻訳文と，評点がより悪い翻訳文を区別する点にある．評点が B になるか C になるかは，内容語やキーワードの欠落度合いに依存して決まる．

上記の基本となる 4 つの拡張版以外の拡張版は，これら 4 つの拡張方針：(1) 交換操作を新たに使うか否か，(2) 標準形の代りに意味コードを参照するか否か，(3) 編集単位を内容語に限定するか否か，(4) 編集単位をキーワードに限定するか否か，の組み合わせに対応する．即ち，拡張版の総数は全部で 15 通り $(= 2^4 - 1)$ である．

訓練事例から学習した決定木は，先に前節で述べたように，評点付与に関する規則に対応する．規則の条件部は，編集距離 ED やその拡張版の値の組み合わせで表される．RED は，基本と拡張版の編集距離合わせて 16 種類の編集距離をまとめ上げた決定木を用いて評点を自動的に付与する．

は内容語や助詞の選択誤りに比べれば少く，交換を許す利点の方が大きいと考え，本稿では交換を導入した．

[10] 意味解析を使える場合には，意味解析の結果わかった各形態素の意味コード同士を照合する．

4 評価実験

新自動評価法 RED を評価するために，次の 2 つのタスクにおける性能を，RED と BLEU(1 章で概説した代表的な自動評価法) とで比較した．

(I) 翻訳文に 4 段階評点 A〜D[1]の 1 つを付与する．

(II) テストコーパス (テスト文集合) 上で MT システムの翻訳性能を比較し，MT システム同士の優劣を判定する．

実験 (I) における自動評価法の評価指標は，評点付与の正解率，即ち自動評価法によって付与された評点が複数評価者による評点のメジアンに一致した割合[11] である．実験 (II) における自動評価法の評価指標は優劣判定の正解率，即ち自動評価法による優劣判定が人手評価による優劣判定と一致した割合である．

評価値はテストセットや参照訳の取り方に依存して変化するため，評価値の平均挙動を比較する．そのために，テストコーパス上で 10 重交差検定 [Mitchell 97] を行い，テストセット用の各部分集合 (held-out data) に対して，予め準備した後述の 16 個の参照訳から一定数の参照訳の部分集合 5 つ[12]をランダムに選んだ．そして，全試行の平均を比較した．参照訳数が増すにつれ，RED と BLEU の性能がどう変化するかを調査するために，ランダムに選ぶ 5 つの参照訳の集合を色々なサイズで試した．

4.1 実験条件

本節では，実験で用いたリソースについて述べる．具体的には，テストコーパス，MT システム，人手評点の結果，および参照訳である．

4.1.1 テストコーパス

テストコーパスは，旅行会話基本表現集，BTEC(Basic Travel Expression Corpus)[Takezawa 02] からランダムに選んだ 345 個の日本語と英語の対訳データである．BTEC は，海外旅行の色々な場面で利用される種々の会話表現を網羅的に集めたコーパスである．

[11]評点付与の誤りは正しい評点との段差 (一段ズレ，二段ズレ等) に応じて段差が大きいほどコストを課すようなに誤りコスト行列 (Misclassification Cost Matrix[Weiss 91]) を設定するのが理想的であるが，一段ズレと二段ズレの間でどうコストの差を設けるべきかは自明ではないため，本稿では簡便に全ての誤りのコストを等しく 1 とした．

[12]5 という数字には特に意味はない．交差検定同様，重要な点は色々な参照訳の集合で評価することである．

表 3: 実験に利用した MT システムの翻訳性能：この表は，実験に利用したテストコーパス (4.1.1 節) 上において人手評価 (4.1.3 節) を行なった際，評点がある評点以上であった翻訳文の割合を示す．例えば，第 3 列は評点が B 以上の翻訳文の割合を示しており，同割合が 43.1%の MT システムから 78.8%の MT システムまであったことを示す．

評点範囲	A	B 以上	C 以上
割合 (%)	30.7 - 64.9	43.1 - 78.8	57.6 - 89.0

4.1.2 評価対象の MT システム

実験に用いた評価対象の MT システムは，9 つの英日 MT システムである．内訳は，3 種類の MT システムの 3 世代，TDMT (0103, 0110, 0203)[13], HPAT (0110, 0204, 0209), および SAT (0110, 0204, 0209) である．3 種類の MT システム，TDMT，HPAT，SAT を概説すると，

- **TDMT**(Transfer Driven Machine Translation) は，人手で記述した構造変換ルールを用いたパターンに基づく MT システム [古瀬 99] である．
- **HPAT** (Hierarchical Phrase Alignment based Translation) は，対訳事例から自動的に学習された構造変換ルールを用いたパターンに基づく MT システム [Imamura 02, 今村 04] である．
- **SAT** (Statistical ATR Translator) は，句やチャンクなどのまとまりを翻訳単位とした統計的 MT システム [渡辺 03, 渡辺 04] である．

表 3 にあるように，これら MT システムの性能は多様である．

4.1.3 人手評価結果

人手評価を準備するために，英語が堪能な日本人 9 名個別に ATR の翻訳文評価基準[1] に従って翻訳文を評価して貰った．評価に際しては，評価の揺れ[14] を防ぐために，入力文毎の翻訳文 9 つを同時に提示し，評点を付与して貰った．各翻訳文には評点のメジアンを最終的に付与した．メジアンは評価者の選び方によらず揺れが小さいことを検証した [秋葉 05, Akiba 03].

4.1.4 参照訳

参照訳を準備するために，著者は日本人の翻訳家 5 名個々にテストコーパスの英語を 3 通りに翻訳して貰った．準備した参照訳数は，テストコーパス中の対訳と合わせて 16 個

[13] '0103' などの数字は，そのシステムのリリース時期を示す．例えば，'0103' は 2001 年 3 月にリリースされたシステムであることを示す．

[14] 評価者に評価対象の翻訳文を別々に提示すると，同じ翻訳文 (異なる MT システムが同じ結果を出すことがある) に対して異なる翻訳結果を付与することがある．それをここでは評価の揺れと呼ぶ．同時に提示することにより，翻訳文の相対的な優劣は保証される．

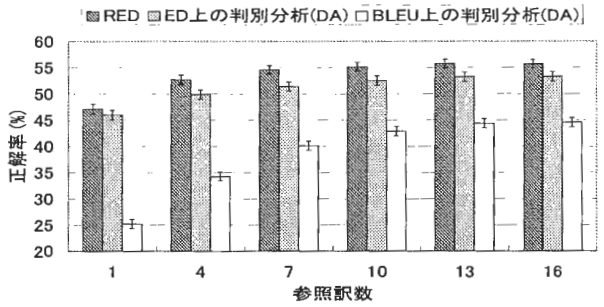

図 3: 10 重交差検定による性能比較．RED の性能 v.s. ED または BLEU による判別分析による自動評価の性能

である．

4.2 実験結果 (I)

本節では，本章のはじめで述べた実験 (I) の結果を示す．この実験では，BLEU を用いて翻訳文に評点を付与するために，BLEU が各翻訳文に付与するスコア上で判別分析 (**DA**, discriminant analysis) を行い，付与する評点を決めた．同スコアは，1 章で説明した各翻訳文の N グラム正解率に対応する．更に比較のために，BLEU の代わりに ED を用いた実験も同様に行った．

図 3 に，RED による評点付与の正解率と ED または BLEU のスコア上での判別分析による評点付与の正解率を示す．横軸は参照訳の数を示し，エラーバーは標準誤差を示す．以下に示す信頼度は RED の正解率の信頼区間と，判別分析による正解率の信頼区間が重ならないように計算された．

RED は，用いた参照訳がどの数においても，BLEU 上の判別分析と比べ 99%の信頼度で有意に優っていた．また RED は，参照訳が 4 から 10 の時，ED 上の判別分析と比べ 87%の信頼度で有意に優っていた．RED の正解率と ED 上の判別分析の正解率の差は RED の正解率と BLEU 上の判別分析の正解率の差に比べると小さいが，統計的に有意な差であった．利用する参照訳の数が増加すればするほど，RED の正解率はより良くなった．参照訳の数が 4 または 7 の時の RED の正解率は，参照訳の数がより多い場合の正解率にほぼ達している．

実験 (I) では個々の翻訳文を評価するため，BLEU に課されたタスクは BLEU 本来の評価タスクとは異なり，BLEU に若干不利であるかもしれない．そこで著者は，BLEU 本来の評価タスクであるシステム性能評価を行なうという実験 (II) の結果を次節で示す．即ち，RED をシステム性能評価に適用した結果を示す．

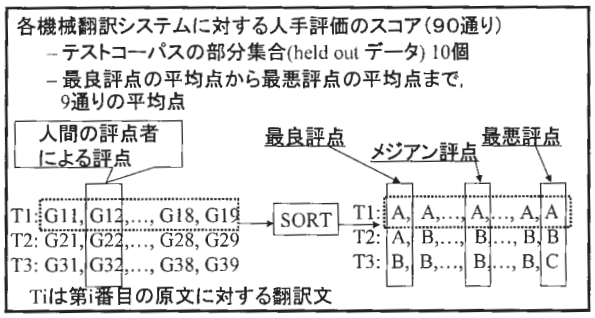

図 4: 人間による MT システムの点数付け

4.3 実験結果 (II)

本節では，本章のはじめで述べた実験 (II) の結果を示す．実験 (II) では，人間による性能評価および RED による性能評価を算出するために，評点 A,B,C, および D を順に 4,3,2, および 1 に対応付け，各テスト部分集合 (held-out データ) 上での対応する数値の平均点を求めた．以下，これらの平均値をおのおの人手評価のスコアおよび RED のスコアと呼ぶ．

メジアン評点の分散は非常に小さいが，ゼロではない．メジアン評点の若干ある分散を考慮するために，人間の性能評価によるシステムの優劣は単にスコアの差を比較するのではなく，統計検定により有意な差があるかで決定した．検定には，ノンパラメトリック多重検定[15] [Hochberg 83, 田中 99] を用いた．

具体的には，以下の手順で人間による性能評価を算出した (図 4 参照)．(i) 各翻訳文の評点を高いものから順に低いものへと並べ直す．(ii) “各翻訳文の i 番目に高い評点を集めて平均値を算出する” という操作を 1 番目から 9 番目まで繰り返し，高い評価値の平均値から低い評価値の平均値まで全部で 9 個の平均値を計算する．これらの平均値を各テスト部分集合 (held-out データ) に対して計算する．(iii) (ii) で計算した全試行におけるスコアの差を前述の多重検定により 95%の信頼度で検定した．

自動評価法 RED および BLEU についても同様で，各自動評価法毎に任意の 2 つの MT システムのスコアの差を前述の多重検定で検定した．この検定結果による MT システムの優劣判定が，人手評価による優劣判定と一致する割合 (一致率) を調べた．

表 4 に，上記優劣判定の一致率を示す．また，一致しない割合 (不一致率) を誤り種別毎に集計した結果も合わせて，表 4 に示す．表の各列は，参照訳の数が列の頭に記載の数字であった場合における優劣判定の一致率および不一致率である．RED は，BLEU と比べ，人手評価による優劣判定との一致率が格段に高い．RED では一致率が 90%台であったが，BLEU では 30%-60%と低調であった．RED や BLEU の一致率は参照訳数の増加と共に単調に改善されるとは限らないが，複数の参照訳を利用した場合の一致率は，単一の参照訳を利用した場合より高い．参照訳を準備する労力および効率を考慮すると，4 個程

[15] よく知られているように，単純な t 検定などの 2 種の標本データを検定する方法を繰り返すと，検定の信頼度は指数関数的に激減する．多重検定は，信頼度を保ったまま，複数の標本データを検定することが可能な検定手法として知られる．本稿では，標本の確率分布を仮定せずに適用ができるノンパラメトリック多重検定を適用した．具体的には，Kruskal-Wallis 検定と Tukey-Kramer 検定を用いた．

表 4: 人間の評点者による平均評価と自動評価の一致率および不一致率：X の値は，人間の平均評価に基づく多重検定の結果を示す．Y の値は，RED または BLEU による自動評価に基づく結果を示す．X が0の場合，ある MT システムが別の MT システムより有意に性能が優っていたことを示す．X が1の場合，ある MT システムが別の MT システムと性能に関して有意な差が無かったことを示す．X が2の場合，ある MT システムが別の MT システムより有意に性能が劣っていたことを示す．Y の値の定義は X の値と同様である．

		RED					BLEU				
		参照訳の数					参照訳の数				
	XY	1	4	7	10	13	1	4	7	10	13
一致率	00, 11, or 22	91.7	94.4	89.0	91.7	91.7	30.6	30.6	30.6	30.6	30.6
不一致率	01 or 21	0	2.8	5.5	2.8	2.8	41.7	47.2	38.9	44.4	44.4
	10 or 12	8.3	2.8	5.5	5.5	5.5	8.3	8.3	8.3	8.3	8.3
	02 or 20	0	0	0	0	0	19.4	13.9	22.2	16.7	16.7

表 5: 自動評価が ED 上の判別分析 (DA) または ED である場合の，表 4 に対応する結果

		ED 上の DA					ED				
		参照訳の数					参照訳の数				
	XY	1	4	7	10	13	1	4	7	10	13
一致率	00, 11, or 22	80.5	86.1	86.1	83.3	80.5	25.0	25.0	25.0	27.8	27.8
不一致率	01 or 21	16.7	11.1	11.1	11.1	13.9	36.1	33.3	33.3	30.5	30.5
	10 or 12	2.8	2.8	2.8	5.6	5.6	11.1	13.9	13.9	13.9	13.9
	02 or 20	0	0	0	0	0	27.8	27.8	27.8	27.8	27.8

度の参照訳が妥当であると考える．また，RED には致命的な間違いは起きていなかったが，BLEU には起きていた．ここで致命的な間違いとは，MT システム甲と乙を比べた際に，人手評価は甲が優っていると評価している場合に，自動評価法が逆に乙が優っていると判定するという判定誤り (表 4 の “02 または 20”) である．

次に，RED の高い一致率の要因を分析するために，ED のみを使った場合および，ED のみと教師データを組み合せた場合 (実験 (I) で用いた ED 上の判別分析 (DA)) についても，BLEU と RED の場合と同様に，人手評価との一致率および不一致率を調べた．その結果を表 5 に示す．一致率は，RED，ED 上の DA，ED の順位高い．一致率の差を見ると，‘RED’ と ‘ED 上の DA’ との差は 55%程度であり，‘ED 上の DA’ と ‘ED’ との差は 10%程度である．不一致率の誤り種別毎に集計を ‘ED 上の DA’ と ‘ED’ とで比べると，何れの誤りタイプにおいても値が減少している．同様に，不一致率の誤り種別毎に集計を ‘RED’ と ‘ED 上の DA’ とで比べると，“01 or 21” における不一致率が削減幅が大きく，性能差の検出能力が改善された．

以上を鑑みると，RED の高い一致率は，RED が用いる教師データによるところが大きい．RED の現実的な利用方法は，大規模な評価を行なう際の評価作業の支援であろう．即

ち，1) 現実的な規模の人手評価を行って教師データを作成し，2)RED で評点付与規則を学習し，3) 学習した規則に基づき大規模な学習を実施する．ここで強調したい点は，BLEU や ED は教師データを用いないため，RED に比べ人手評価との一致率が非常に低い点である．

また，RED のこの高い一致率は，複数編集距離 (3.2 節)，即ち，ED，その 4 つの基本拡張版，ED_{swp}, ED_{sem}, ED_{cnt},ED_{key}，およびこれらの組み合わせ，の導入にも起因する．一方，BLEU の場合，RED に導入した柔軟なマッチング機能はなく，単に N グラム正解率を考慮しているに過ぎない．その結果例えば，助詞が誤りであったり抜けている場合には，内容語が正しく翻訳されている場合でも，バイグラム正解率が低くなる．このような場合，BLEU は，人手評価より MT システムを低く評価する．

RED では 16 種類の編集距離を用いて自動評価を行なうが，著者はいずれの編集距離も自動評価に重要であると考える．3 章で述べたように，入力文に対する翻訳文は 1 通りとは限らず，数多くの翻訳文が等しく容認される．あらゆる可能な参照訳を全て準備することがもし可能であるならば，通常の編集距離 (ED) だけで十分自動評価は可能であろうが，現実には可能な参照訳を網羅的に準備することは困難であり，ED だけでは不十分である．基本となる 4 つの拡張版編集距離，ED_{swp}，ED_{sem}，ED_{cnt}，および ED_{key} の重要性は，3.2 節で述べた通りである．4 種類の基本拡張を組み合せた拡張版編集距離が重要な理由は，翻訳文の文構造と訳語の多様性が複合的に起こり得る点にある．翻訳文の文構造，または翻訳文中の訳語選択の組み合せを網羅的に記述することは難しいため，上述の複合的な多様性のあらゆる可能性に対処するには，基本拡張の全ての組み合せを考えることが重要である．

最後に，RED はどのような場合に人手評価と食い違うかを議論する．RED は，翻訳文から参照訳への編集距離に基づいて翻訳文を評価するため，翻訳文が正しくとも，その文構造が予め準備した参照訳の文構造と大きく異なる場合，RED による翻訳文の評価は低くなり人間評価と評価がずれ

る．RED で対処できる文構造の違いは，隣接する形態素の入れ換えのみであり，それ以外の文構造の違いは対応できていない．従って，食い違いを極力少なくするためには，参照訳を予め準備する際に，よく使われる文構造をできる限り多く準備する必要がある．

5 関連研究

本章では，本稿の関連研究について議論する．

編集距離に基づくその他の自動評価法としては，WER[Su 92, Jurafsky 00] および mWER [Thompson 91, Niessen 00] がある．

著者が知る限り，翻訳文を自動的に評価する初めての手法は，WER[Su 92] である．[Su 92] では参照訳の数は 1 つだけで，翻訳文を翻訳家が編集し，編集結果を本稿の参照訳として用いた．最近では WER で用いる参照訳は，予め 1 つ準備する．そのため WER は，完璧な翻訳文ではあるが予め準備した参照訳と比べ語彙的にまたは構造的に異なる翻訳文と，誤訳とを区別できない．この WER の問題点を解決させるために，参照訳を複数利用す

るmWERが導入された[Thompson 91, Niessen 00]．このmWERは本稿のEDと等価である．

REDでは，mWERとは違い，EDの各種拡張版を導入し，決定木としてまとめ上げた．図3が示すように，REDは，単一の編集距離EDの値を用いた判別分析と比べ，性能が有意に優っていた．従って，REDはEDと等価なmWERより十分に性能が優っていると言える．

BLEU以外のNグラムに基づく自動評価法には，NIST [Doddington 02] がある．NISTはBLEUを派生させた自動評価法であり，その主たる違いは，Nグラム正解率の平均値の取り方，および参照訳より短い翻訳文に課すパナルティのかけ方である．NISTは，本稿で取り上げているBLEUの欠点，即ち，誤り評価の出現箇所依存性，および誤りに対する寛容さの欠如を継承している．

また，BLEUに類似したその他の自動評価法に，GTM [Melamed 03]がある．GTMでは，BLEUのNグラムの代りに，翻訳文と参照訳との共通単語を覆う極大部分単語列(maximal substring)を用いて正解率(precision)と再現率(recall)を計算し，これらのF値で翻訳の良さを測る．従ってGTMは，図1の例のように翻訳文中の誤り数がたとえ同じ1つでも，文頭や文末で起きた誤りを軽く評価する．また，GTMは，やはり出現形でのみ単語の比較を行なうため，BLEUと同様，誤りに対する寛容さに欠ける．

6 おわりに

本稿では，翻訳文の品質を評価するという人間の知的能力を自動化する挑戦的で新しい試みを報告した．ここで翻訳文は，機械翻訳(MT)システム，特に会話を翻訳する音声翻訳(SSMT)システムの翻訳部からの出力である．著者は，複数の編集距離を用いて翻訳文に自動的に評点(4段階評点[1]の1つ)を付与する新自動評価法RED[秋葉 05, Akiba 01, Akiba 03]を報告した．REDでは，翻訳文に自動的に評点を付与する決定木を訓練事例から学習し，この決定木を用いて評点が未知の翻訳文に対して評点を付与する．各訓練事例は翻訳文と人手評点の対の集合であり，翻訳文は複数の編集距離を用いてベクトル化される．ここで複数の編集距離は，通常の編集距離(**ED**)およびEDの拡張版である．新たな評価対象の翻訳文は，訓練事例と同様にベクトル化され，学習した決定木を用いて評点を付与される．新自動評価手法REDおよび近年機械翻訳研究者の間でよく使われている自動評価手法BLEUの性能を比較するために，REDとBLEUを用いて，MTシステムを評価する実験を2種類行った．第1の実験では，REDとBLEUで翻訳文に評点を付与する際の正解率を比較した．第2の実験では，REDとBLEUで1つのMTシステムが別のMTシステムより有意に優るか否かを判定させる際の正解率を比較した．両実験を通し，以下の知見を得た．

- 両実験において，REDはBLEUより性能が良いことが示された．
- 特に第2の実験においては，REDでは正解率が90%台と高かった．REDの性能が高い要因は，複数の編集距離を利用する点と，訓練事例を用いて評点を付与する決

定木を学習した点にある．

- BLEU は，SSMT システムを評価するには適していない．

RED を含め本稿で取り上げたすべての自動評価法 [Papineni 01, Doddington 02, Melamed 03, Su 92, Niessen 00] は，参照訳を事前に準備しておくことが前提である．自動評価手法の精度を更に向上させるために，参照訳の自動構築法 [Imamura 03, Finch 03] を含め，今後研究を進める予定である．

謝辞

本研究は総務省の研究委託「携帯電話等を用いた多言語自動翻訳システム」により実施したものである．著者は，類語新辞典 [大野 81] の使用許可を頂いた (株) 角川書店に感謝します．

参考文献

[秋葉 05] 秋葉 泰弘, 今村 賢治, 隅田 英一郎, 中岩 浩巳, 山本 誠一, 奥乃 博：複数の編集距離を用いた口語翻訳文の自動評価, 人工知能学会論文誌, Vol. 20, No. 3, pp. 139–148 (2005)

[秋葉 98] 秋葉 泰弘, Almuallim, H., 金田 重郎：例からの学習技術の応用に向けて — 2．応用上の課題に対する解決法, 情報処理, Vol. 39, No. 3, pp. 245–251 (1998)

[Akiba 01] Akiba, Y., Imamura, K., and Sumita, E.: Using Multiple Edit Distances to Automatically Rank Machine Translation Output, in *Proc. MT Summit VIII*, pp. 15–20 (2001)

[Akiba 03] Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S., and Okuno, H. G.: Experimental Comparison of MT Evaluation Methods: RED vs. BLEU, in *Proc. MT Summit IX*, pp. 1–8 (2003)

[Doddington 02] Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, in *Proc. HLT-02*, pp. 257–258 (2002)

[Finch 03] Finch, A., Watanabe, T., and Sumita, E.: Data-Oriented Paraphrasing, in *Proc. RANLP-2003*, pp. 153–157 (2003)

[古瀬 99] 古瀬 蔵, 山本 和英, 山田 節夫：構成素境界解析を用いた多言語話し言葉翻訳, 自然言語処理, Vol. 6, No. 5, pp. 63–91 (1999)

[Hochberg 83] Hochberg, Y. and Tamhane, A. C.: *Multiple Comparison Procedures*, Wiley (1983)

[Imamura 02] Imamura, K.: Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-Based MT, in *Proc. TMI-02*, pp. 74–84 (2002)

[Imamura 03] Imamura, K., Akiba, Y., and Sumita, E.: Automatic Expansion of Equivalent Sentence Set Based on Syntactic Substitution, in *Proc. HLT-NAACL-03*, pp. 37–39 (2003)

[今村 04] 今村 賢治, 隅田 英一郎, 松本 裕治：直訳性を利用した機械翻訳知識の自動構築, 自然言語処理, Vol. 11, No. 2, pp. 85–99 (2004)

[Jurafsky 00] Jurafsky, D. and Martin, J. H.: *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, Inc (2000)

[King 96] King, M.: Evaluating Natural Language Processing Systems, *Communications of the ACM*, Vol. 39, No. 1, pp. 73–79 (1996)

[Kitano 94] Kitano, H.: *Speech-to-Speech Translation*, Kluwer Academic Publishers (1994)

[国研 55] 国研 (編)：談話語の実態, Technical Report 国立国語研究所報告:8, 国立国語研究所 (1955)

[Lowrance 75] Lowrance, R. and Wagner, R. A.: An Extension of the String-to-String Correction Problem, *J. ACM*, Vol. 22, No. 2, pp. 177–183 (1975)

[Marcu 02] Marcu, D. and Wong, W.: A Phrase-Based, Joint Probability Model for Statistical Machine Translation, in *Proc. ACL-02 WS EMNLP*, pp. 133–139 (2002)

[Melamed 03] Melamed, I. D., Green, R., and Turian, J. P.: Precision and Recall of Machine Translation, in *Proc. HLT-NAACL-03, short papers*, pp. 61–63 (2003)

[Mitchell 97] Mitchell, T. M.: *Machine Learning*, The McGraw-Hill Companies Inc. (1997)

[Niessen 00] Niessen, S., Och, F. J., Leusch, G., and Ney, H.: An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research, in *Proc. LREC-00*, pp. 39 45 (2000)

[Och 02] Och, F. J. and Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, in *Proc. ACL-02*, pp. 295–302 (2002)

[大野 81] 大野 晋, 浜西 正人：類語新辞典, 角川書店 (1981)

[Papineni 01] Papineni, K. A., Roukos, S., Ward, T., and Zhu, W. J.: Bleu: a Method for Automatic Evaluation of Machine Translation, in *Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY*, pp. 257–258 (2001)

[キンラン 95] キンラン J.R. 著, 古川 康一監訳：AI によるデータ解析, トッパン (1995)

[Su 92] Su, K. Y., Wu, M. W., and Chang, J. S.: A New Quantitative Quality Measure for Machine Translation Systems, in *Proc. COLING-92*, pp. 433 439 (1992)

[Sumita 99] Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K., and Shirai, S.: Solutions to Problems Inherent in Spoken-Language Translation: the ATR-MATRIX Approach, in *Proc. MT Summit VII*, pp. 229–235 (1999)

[Takezawa 02] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S.: Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, in *Proc. LREC-02*, pp. 147–152 (2002)

[田中 99] 田中 豊, 垂水 共之 （編）：統計解析ハンドブック---ノンパラメトリック法, 共立出版 (1999)

[Thompson 91] Thompson, H. S.: Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment, in *Proc. Evaluator's Forum-91*, pp. 215- 223 (1991)

[Wagner 74] Wagner, R. A. and Fischer, M. J.: The String-to-String Correction Problem, *J. ACM*, Vol. 21, No. 1, pp. 168–173 (1974)

[渡辺 03] 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, pp. 3202–3210 (2003)

[渡辺 04] 渡辺 太郎, 今村 賢治, 隅田 英一郎, 奥乃 博：階層的句アライメントを用いた統計的機械翻訳, 電子情報通信学会論文誌, Vol. J87-D-II, No. 4, pp. 978–986 (2004)

[Weiss 91] Weiss, S. M. and Kulikowski, C. A.: *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*, Morgan Kaufmann (1991)

[Yamada 02] Yamada, K. and Knight, K.: A Decoder for Syntax-Based Statistical MT, in *Proc. ACL-02*, pp. 303–310 (2002)

[Zens 03] Zens, R. and Ney, H.: A Comparative Study on Reordering Constraints in Statistical Machine Translation, in *Proc. ACL-03*, pp. 144–151 (2003)