Internal Use Only（非公開）

ＴＲ－ＳＬＴ－００８４

統計翻訳の性能と学習コーパスの関係に関する調査

On Relations between MT Performance and Training Corpora

増野　成章

MASUNO Naruaki

2005 年 1 月 11 日

概要

コーパスをベースとした機械翻訳方式の一つである統計翻訳において、対訳コーパスの学習量と翻訳精度の関係について調査した。また、自動音声翻訳において音声認識モジュールと機械翻訳モジュールの最適な統合方法を考える予備調査として、学習コーパスの情報を変化させた場合にそれが翻訳精度にどのような影響を与えるか、という調査を行った。

（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

目 次

1 はじめに

コーパスをベースとした機械翻訳方式の1つに統計翻訳がある。この統計翻訳においては、用意された対訳コーパスから翻訳モデルや言語モデルについて学習を行う。したがって、学習させるコーパス量は翻訳結果の精度に大きな影響を与えていると考えられる。しかし、今回の実験で使用するSAT（Statistical ATR Translator）の翻訳精度と学習コーパス量との関係は明らかになっていない。そこで今回の調査では、学習させるコーパスの量を変化させることで、翻訳結果がどのように変化するか実験を行い、翻訳精度と学習コーパス量との関係を明らかにしたい。

また、自動音声翻訳の1つのモジュールとしても機械翻訳は利用されている。音声翻訳の統合技術として、音声認識モジュールと機械翻訳モジュールの最適な統合方法については様々な研究がなされている。その研究の1つとして、音声認識部と機械翻訳部での機能分担をどのように行えばよいか、というものがある。「多義解消」を例にとってみると、現在の音声翻訳システムでは、音声認識部で言語モデルを用いて解析を行い語の特定を行っている。そこで、この多義解消のタスクを機械翻訳部に振り分けることを考える。具体的には、音声認識部での出力、つまり機械翻訳部の入力を、各単語の漢字かな表記を捨象した形式にすると結果として多義解消は機械翻訳部のタスクと考えることができる。今回の調査では、音声認識結果に付与されている情報を変更することで、音声翻訳全体の精度がどのように変化するか実験を行う。

以降、2章では今回行った実験である「学習コーパス量と翻訳性能の関係」「学習コーパスにおける単語属性の詳細度と翻訳性能の関係」の各々について、内容および方法について述べる。3章、4章では2章で述べた方法によって行った実験の結果および考察を示す。5章で今回の調査のまとめを行う。

2 調査の内容と方法

今回の調査では、「学習コーパス量と翻訳性能の関係」「学習コーパスにおける単語属性の詳細度と翻訳性能の関係」の2つの項目について調査を行った。以下に、各々についての内容および方法について示す。

2.1 学習コーパス量と翻訳性能の関係

統計的機械翻訳では、翻訳元言語のテキストSから翻訳先言語のテキストTへの翻訳を行うときには次のようなモデルを考える。翻訳元の文は翻訳先のテキストSが雑音のある通信路によって翻訳先テキストTに変換されたものと考え、「翻訳」という行為はSからTへの復号化であると考える。したがって、良い翻訳を行うには復号化の際の誤りを最小にすればよいから、以下の式を満たすテキストTを求めることになる。

$$\begin{aligned} \hat{T} &= \mathrm{argmax} P(T|S) \\ &= \mathrm{argmax} P(S|T)P(T) \end{aligned} \tag{1}$$

確率$P(S|T)$を与えるモデルを翻訳モデル、$P(T)$を与えるモデルを言語モデルと呼ぶ。したがって、翻訳モデルおよび言語モデルを推定できれば、式(1)を満たすテキストTを決定すればよい。

今回の調査で使用する翻訳システムは、SAT（Statistical ATR Translator）である。このシステムは、用意された対訳コーパスから、翻訳モデルおよび言語モデルについて学習を行う。したがって、翻訳システムの出力の精度は、この用意されたコーパスに依存していると考えられる。しかし、現在のシステムでは、学習に用いるコーパスのシステムに対する影響度というものは明らかになっていない。そこで今回の実験では、対訳コーパスの量を20%、40%、60%、80%、100%と変化させることによって、翻訳結果の精度に対してどのような影響を及ぼすか、ということについて実験を行う。

また、この翻訳システムではコーパスから翻訳モデルと言語モデルを学習しているわけだが、翻訳モデルの学習には対訳コーパスを必要とするのに対して、言語モデルの学習には翻訳先言語のコーパスを用意すればよい。したがって、翻訳モデルの学習に必要な対訳コーパスの量が少なく済むのであれば、それだけコ

ストを削減できる。そこで、「言語モデル」については元のコーパス量で学習させたものを用意し、「翻訳モデル」だけをコーパス量を変化させたもので学習させた場合に、翻訳結果に対してどのような影響があるかという実験も行う。

2.2 学習コーパスにおける単語属性の詳細度と翻訳性能の関係

音声翻訳システムでは、機械翻訳システムに相当するモジュールの前に音声認識モジュールがあり、これらのモジュールを接続することでシステムが構成される。しかし、このような音声翻訳システムにおいては、通常の機械翻訳システムには見られなかった次のような問題が生じる。

まず、音声認識モジュールの精度の問題である。これは、音声認識モジュールが必ずしも正しい解（入力）を与えるとは限らない点である。テキスト翻訳の場合においても必ずしも入力に誤りがないとは言えないが、音声認識の比ではない。次に、同音語の問題がある。音声入力では「橋」や「箸」や「端」などは皆「ハシ」という入力であり、区別をつけることが出来ない。このため、テキスト入力より曖昧性が高くなってしまう。アクセントやピッチの高低などを利用し、区別することなども考えられているが、現時点では実用には至っていない。

機械翻訳部の入力は、音声認識部で認識された結果に対し、言語モデル（n-gram）を用いてすでに解析がなされている。この解析で、各単語（形態素）に対し、以下のように表層形や読み、品詞、活用などの情報が与えられている。

```
御|オ|御|接頭辞|||| つまみ|ツマミ|つまみ|普通名詞|||| は|ハ|は|係助詞|||
あり|アリ|ある|本動詞|五段ラ|連用|| ます|マス|ます|助動詞|特殊サ|基本||
か|カ|か|終助詞|||
```

そこで、「各単語に対して、品詞や活用といった情報を決定するタスク」を機械翻訳部に移してやることを考える。つまり、音声認識が行われた時点では品詞などの同定をせずに機械翻訳を行う、というものである。

具体的には、音声認識部の出力の各単語ごとに付与されている属性について次のような形式の変更を行う。

- 「読み」情報のみを残す
- 「読み、品詞」情報のみを残す
- 「読み、正規表現」情報のみを残す

「読み、品詞」の情報のみを残したコーパスで学習・翻訳を行った場合、多義解消は外部的には行われずに処理されたことになる。(「ハシ」は名詞で橋、端、箸などがあるが、品詞情報だけではどの単語かを特定することは出来ない)。従って、多義解消のプロセスは音声認識部のタスクではなく、機械翻訳部の隠れたタスクとなっている。そこで、上記のように形式を変更した場合と形式の変更を行わなかった結果とを比較し、翻訳精度がどのような影響を受けるかということについて実験を行う。

2.3 評価尺度

本稿では、機械翻訳された結果の評価を行うために、BLEU[3], NIST[4],Word Error Rate(WER),Position independent word Error Rate(PER) を用いる。これらの評価方式では、翻訳結果と模範訳（人手で用意した正解例）との類似度を測定し、翻訳の精度の数値化を行う。このような自動評価法を用いることで、人間が翻訳精度を測定した場合に起こる、各々の判断基準が曖昧である、結果が一定しない、などの問題を避けることができる。

2.3.1 BLEU スコア

BLEU では、翻訳結果と模範訳との類似度を両者の n-gram の一致数を用いて以下の式で算出する。

$$\text{BLEU} = \text{BP}_{\text{BLEU}} \cdot exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{2}$$

ここで、

$$p_n = \frac{\sum_i \text{翻訳文}\ i\ \text{と模範訳}\ i\ \text{で一致した n-gram の数}}{\sum_i \text{翻訳文}\ i\ \text{中の全 n-gram 数}}$$

$$w_n = \frac{1}{N}$$

としている。

p_n は評価する文全体に対して、翻訳文と模範訳の比較を行い、n-gram の一致率を算出している。これを 1-gram〜N-gram について算出し、それらの幾何平均を求めている。n が大きくなるにつれ、n-gram は翻訳の流暢さを表す指標となるが、今回の実験では N=4 としている。また、BP_{BLEU} では、翻訳文が模範訳より短い場合にペナルティを与えており、翻訳文が模範訳より長い場合には 1 を与える。

BLEU スコアは 0〜1 の値を取り、その値が高いほど良い翻訳文であったと評価する。

2.3.2 NIST スコア

NIST スコアは、BLEU スコアの場合と同様に、翻訳文と模範訳との類似度を n-gram の一致数を基に以下の式で算出している。

$$\text{NIST} = \text{BP}_{\text{NIST}} \cdot \sum_{n=1}^{N} \frac{\sum_i \left(\sum_{w_1 \ldots w_n} \text{Info}_i(w_1 \ldots w_n)\right)}{\sum_i \text{翻訳文}\ i\ \text{中での全 n-gram 数}} \tag{3}$$

また、

$$\text{Info}(w_1 \ldots w_n) = log_2 \frac{\text{評価文中の}\ w_1 \ldots w_{n-1}\ \text{の数}}{\text{評価文中の}\ w_1 \ldots w_n\ \text{の数}}$$

$$w_1 \ldots w_n : \text{翻訳文と模範訳に共通}$$

NIST スコアは 0 以上の実数で表され、値が高いほど良い翻訳であると評価する。今回の実験では N=4 としている。また、BP_{NIST} は BP_{BLEU} と同様に翻訳文が模範訳より短い場合にペナルティが与えられる。BLEU の場合と算出方法は違うが、翻訳文の方が模範訳より長い場合には 1 が与えられる。

BLEU スコアとの違いは、個々の n-gram に対し、Info という情報量による重み付けがなされていることである。一般には機能語列より内容語列の方が情報量が高いため、内容語の正しさを重視する傾向にある。また、高次の n-gram になるにつれコーパス中の存在数が減少し、それにともない Info も減少する傾向にあるため、NIST スコアは、BLEU スコアと比べ、語順の正しさより単語訳の正しさを重視した評価法であると言える。

2.3.3 Word Errpr Rate(WER),Position independent word Error Rate(PER)

Word Error Rate は、音声認識結果の自動評価に標準的に用いられている。これを機械翻訳の自動評価に用いる場合には、翻訳文と模範訳との $\text{DP}_{\text{matching}}$ を行い、その結果を基に以下の式で算出する。

$$WER = \frac{\sum_i (\text{挿入語数}\ i + \text{削除語数}\ i + \text{置換語数}\ i)}{\sum_i \text{模範訳}\ i\ \text{の語数}} \tag{4}$$

式 (4) は、挿入コスト、削除コスト、および置換コストの全てが 1 の時の編集距離を正規化したものである。

WERでは語順を考慮していたが、語順を無視して翻訳文を評価する方法もある。それがPosition independent word Error Rateであり、評価式は以下のようになる。

$$PER = 1 - \frac{\sum_i \text{翻訳文}\, i \cdot \text{模範訳}\, i \text{の間の一致語数}}{\sum_i \text{模範訳}\, i \text{の語数}} \tag{5}$$

PER、PERは値は0〜1を取り、値が低いほど良い翻訳であると評価される。

3 「学習コーパス量と翻訳性能の関係」についての調査

3.1 実験条件

使用するコーパス

今回の調査で使用したコーパスは旅行会話基本表現集である。これは、主に旅行会話において多く見られる表現を集めたもので、約15万文からなる。このコーパス量を20%、40%、60%、80%と変更した場合に翻訳結果の精度がどのように変化するかを調査する。また、20%、40%、60%のコーパス量で実験したものについては、コーパスの選び方を変更し3回実験を行った結果の平均を取ることとする。これは、コーパスの選び方によって翻訳結果に影響が出ることを避けるためである。

また、評価用セットについても旅行会話基本表現集を用いる。これは510文×2（set01,set02）からなる。

翻訳モデル、言語モデル

翻訳モデルの学習には、GIZA++を用いて学習したIBMモデル4の語彙モデルを学習した。言語モデルは、単語3グラムで「good-turingスムージング」をかけたものを利用した。

評価方法

評価用セット510文×2について、第2章で述べた4つの評価方法（BLEU,NIST,WER,PER）を用いた。

3.1.1 実験結果

各コーパス量と、実験結果についての評価を図1〜図3、表1〜3に示す。

3.1.2 考察

set01、02の傾向としては、各指標ともWER、PERの40%以外ではset02の方がスコアが高いという結果になった。また、set01、02ともに学習させるコーパス量が増加するにつれ翻訳精度の向上が見られたが、それらには次のような特徴が見られた。コーパス量を20%、40%、60%と増加させた段階では翻訳精度の向上はそれほどの向上は見られなかったのに対し、コーパス量を80%にした場合に、急激な評価の向上が見られた。

このような結果となった原因として、(1) それなりの翻訳精度を出すためにはある程度の知識量が必要で、60%までのコーパス量ではその知識量に達せず80%のコーパス学習でその知識量が獲得できる、(2) 今回の実験では翻訳の際のパラメータ（語の重みづけなど）を変更せずに実験を行ったため、100%のコーパス量の時の翻訳用にチューニングされており、20%〜60%の翻訳の精度が下がってしまった、などの理由が考えられる。

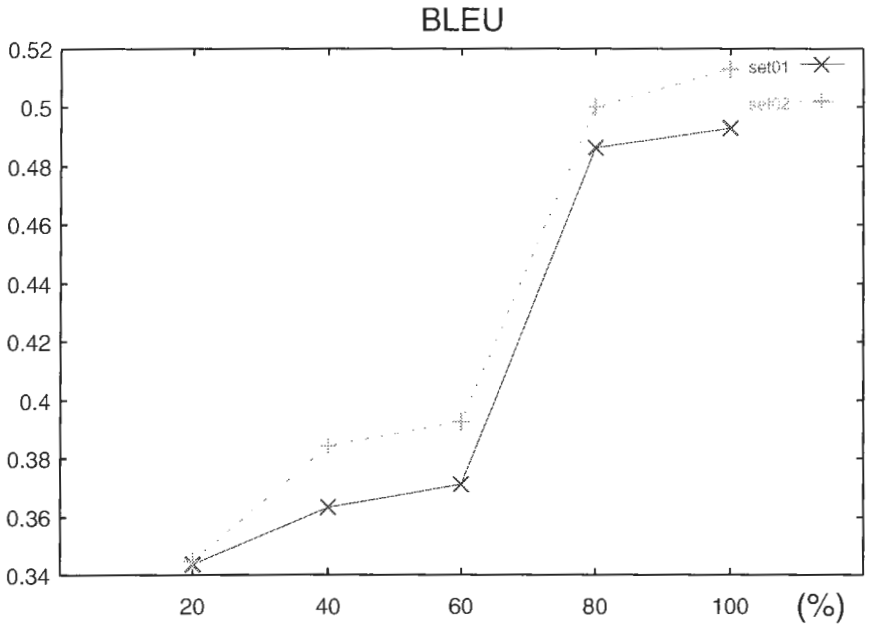

図 1: 実験結果（学習コーパス量と翻訳精度の関係、BLEU 値）

表 1: 実験結果（学習コーパス量と翻訳精度の関係、BLEU 値）

	20%	40%	60%	80%	100%
set01	0.343599	0.363158	0.371163	0.486298	0.492881
set02	0.344529	0.384394	0.392739	0.500278	0.513113

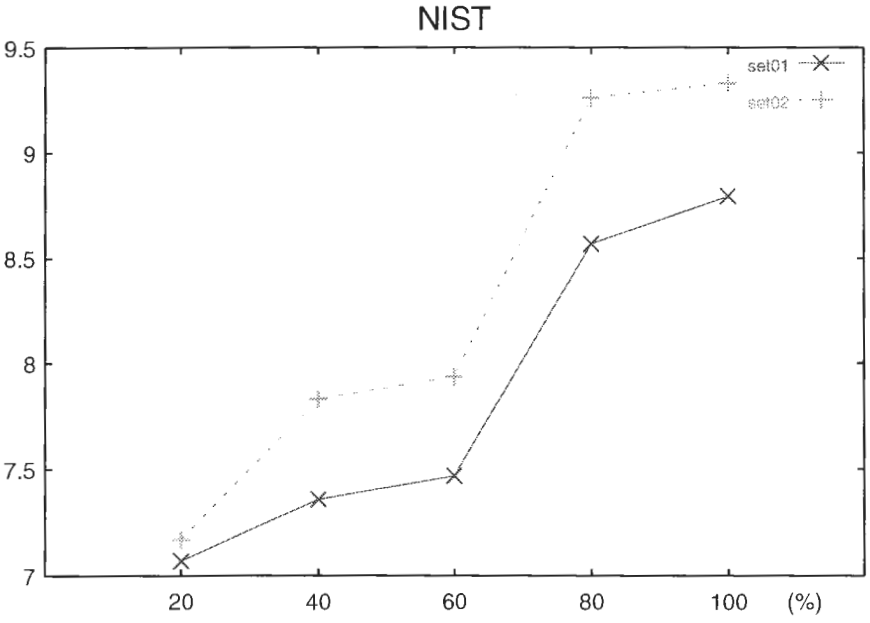

図 2: 実験結果（学習コーパス量と翻訳精度の関係、NIST 値）

表 2: 実験結果（学習コーパス量と翻訳精度の関係、NIST 値）

	20%	40%	60%	80%	100%
set01	7.069397	7.360537	7.470433	8.564920	8.793750
set02	7.169253	7.830263	7.939497	9.260920	9.329900

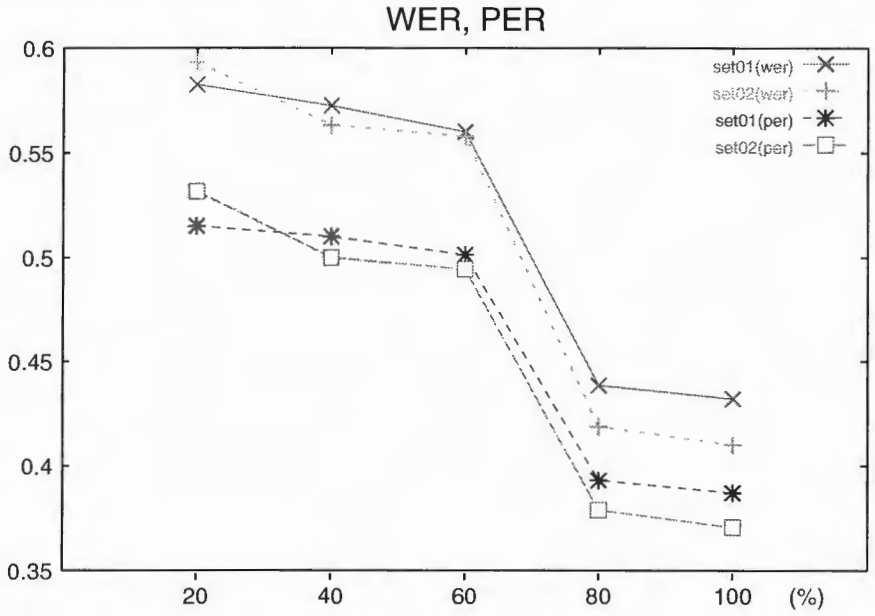

図 3: 実験結果（学習コーパス量と翻訳精度の関係、wer 値、per 値）

表 3: 実験結果（学習コーパス量と翻訳精度の関係、wer 値、per 値）

	20%	40%	60%	80%	100%
set01(wer)	0.582396	0.572302	0.566164	0.438870	0.432114
set02(wer)	0.593312	0.563311	0.557568	0.418931	0.410408
set01(per)	0.514852	0.509924	0.501129	0.393378	0.387097
set02(per)	0.531572	0.499653	0.494178	0.378823	0.370589

3.2 「翻訳モデル」のみを変化させたコーパスで学習させた場合

上記の実験ではコーパス量を変化させて、翻訳モデル、言語モデルを学習させていたが、言語モデルは100%のコーパスで学習させたもので固定し、翻訳モデルのみを変化させた場合の実験を行う。

使用するコーパスは、旅行会話基本表現集約 15 万文、評価用セット 510 文 ×2（set01,set02）、評価方法には BLEU、NIST、WER、PER を用いた。

3.2.1 実験結果

各コーパス量と、実験結果についての評価を図 4〜図 6、表 4〜表 6 に示す。

3.2.2 考察

set01、02 の傾向としては、上記の翻訳モデル、言語モデルの両方を学習させた実験と同じで 02 の方が良い翻訳精度が出ている。また、コーパス学習量と翻訳精度の変化の傾向は、翻訳モデル、言語モデルともに変化させたコーパスで学習させた実験と大きく違っている。

特徴としては、(1)60%のコーパス量で急激に精度が下がっている、(2)BLEU、NIST においては、40%の評価が高い、などが挙げられる。なお、80%コーパスでの BLEU 値が 100%コーパスの BLEU 値より良い評価となっているが、他の 3 つの指標では 100%の方が良い評価である。

単純に考えれば、翻訳に必要な知識が増えれば増えるほど精度は向上するはずである。しかし、学習量が中途半端である場合に、かえって間違った翻訳をしてしまう可能性が考えられる。翻訳を行う際のルールが複雑になった場合に、どのルールを使用したら良いかが判断できなくなるのである。このような理由で60%コーパスの精度が著しく低下した、と考えられる。

また、この実験においては、各コーパスの実験を一度ずつしか行っていないために、例えば 20%、40%、60%のコーパスの選び方の影響を受けている可能性もある。いずれにしても、さらなる検討が必要である。

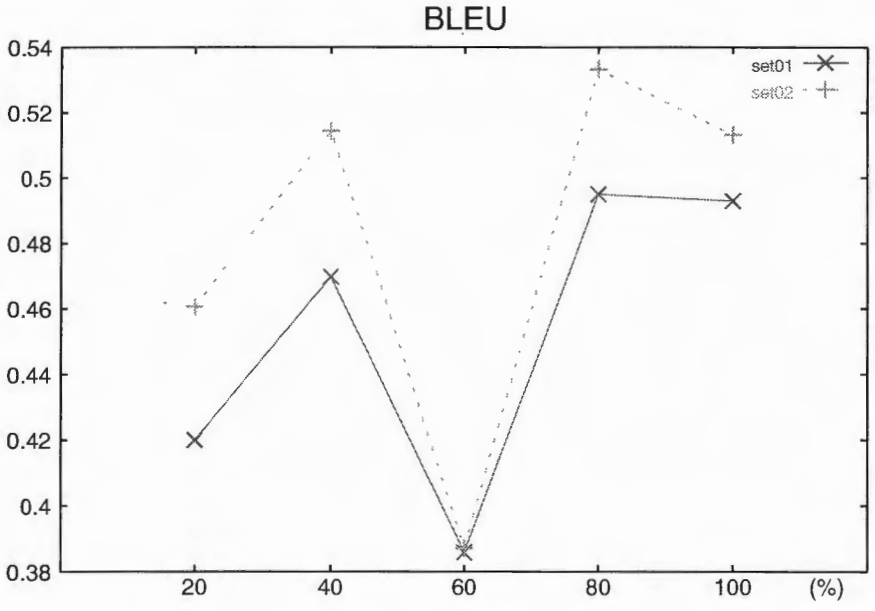

図 4: 実験結果（翻訳モデルのみを変化させたコーパスで学習させた場合、BLEU 値）

表 4: 実験結果（翻訳モデルのみを変化させたコーパスで学習させた場合、BLEU 値）

	20%	40%	60%	80%	100%
set01	0.420036	0.469603	0.385899	0.494881	0.492881
set02	0.460657	0.514545	0.387137	0.533139	0.513113

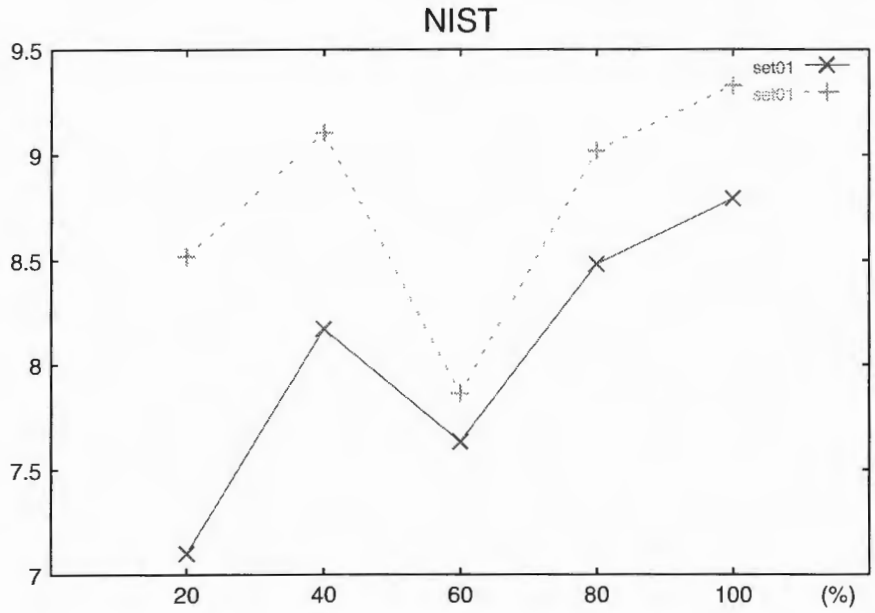

図 5: 実験結果（翻訳モデルのみを変化させたコーパスで学習させた場合、NIST 値）

表 5: 実験結果（翻訳モデルのみを変化させたコーパスで学習させた場合、NIST 値）

	20%	40%	60%	80%	100%
set01	7.10279	8.17201	7.63371	8.47896	8.79375
set02	8.51680	9.10298	7.86422	9.01606	9.32990

表 6: 実験結果（翻訳モデルのみを変化させたコーパスで学習させた場合、WER 値、PER 値）

	20%	40%	60%	80%	100%
set01(WER)	0.501697	0.469184	0.556053	0.461596	0.432114
set02(WER)	0.458376	0.433687	0.550825	0.422668	0.410108
set01(PER)	0.457197	0.422882	0.494340	0.409741	0.387097
set02(PER)	0.423159	0.392181	0.490381	0.382446	0.370589

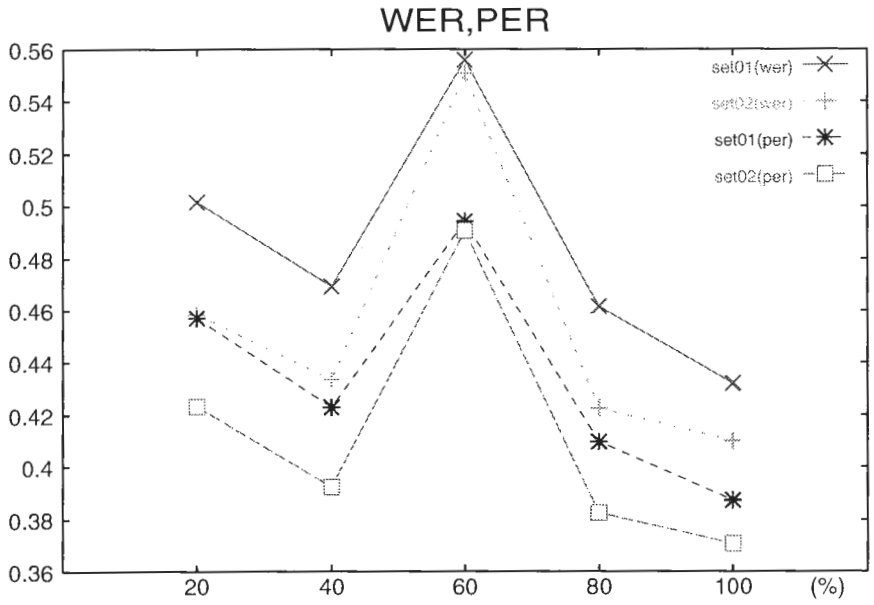

図 6: 実験結果（翻訳モデルのみを変化させたコーパスで学習させた場合、WER 値、PER 値）

4 「学習コーパスにおける単語属性の詳細度と翻訳性能の関係」についての調査

4.1 実験条件

使用するコーパス

この実験で使用するコーパスも上記の実験と同じく、旅行会話基本表現集を用いる。これらのコーパスには、各単語ごとに表層形、読み、品詞、正規形、活用などの情報が付与されている。そこで、「読み」「読み、品詞」「読み、正規」の情報のみが残された3種のコーパスに変更し、各々について学習、翻訳を行う。評価用セットについては、旅行会話基本表現集から抜き出された510文×2（set01、set02）を使用する。

言語モデル、翻訳モデル

翻訳モデルの学習には、GIZA++を用いて学習したIBMモデル4の語彙モデルを学習した。言語モデルは、単語3グラムで「good-turingスムージング」をかけたものを利用した。

評価方法

評価用セット510文×2について、第2章で述べた4つの評価方法（BLEU,NIST,WER,PER）を用いた。

4.1.1 実験結果

各コーパス量と、実験結果についての評価を図7〜図10、表7〜10に示す。

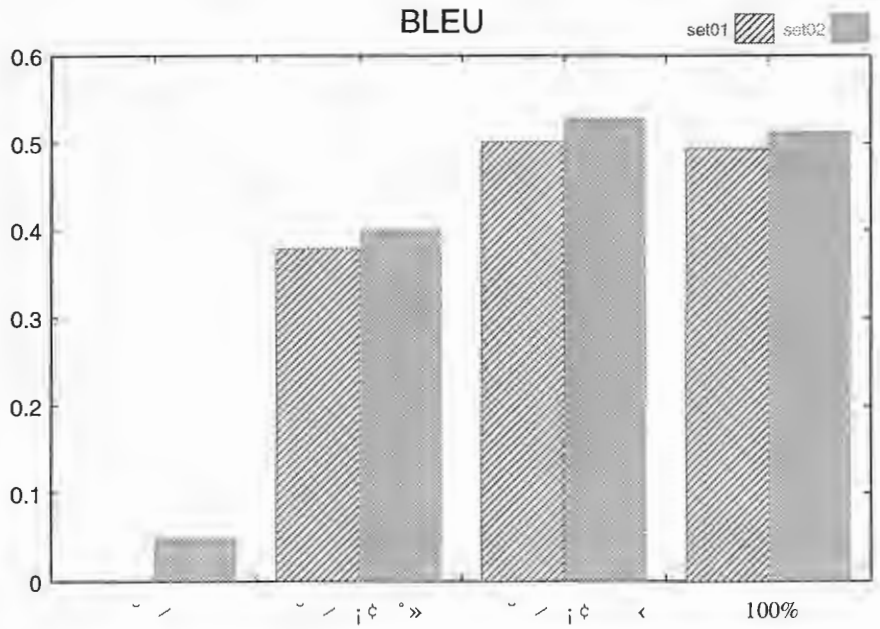

図7: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、BLEU値）

表 7: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、BLEU 値）

	読み	読み、品詞	読み、正規	100%
set01	0	0.377763	0.501131	0.492881
set02	0.049853	0.401496	0.528610	0.513113

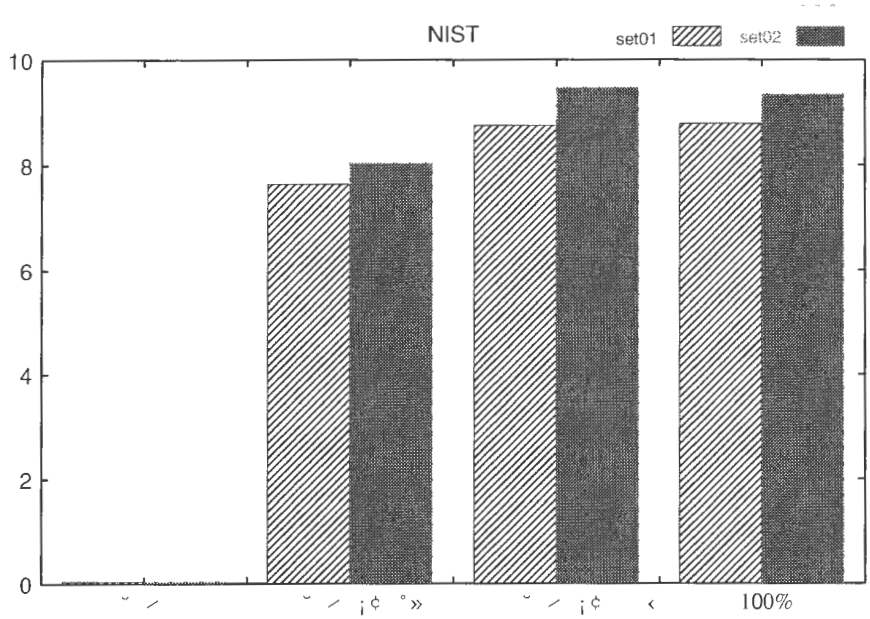

図 8: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、NIST 値）

表 8: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、NIST 値）

	読み	読み、品詞	読み、正規	100%
set01	0.042577	7.62157	8.75707	8.70375
set02	0.040602	8.04023	9.46697	9.32990

表 9: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、WER 値）

	読み	読み、品詞	読み、正規	100%
set01	0.740204	0.548128	0.425992	0.432114
set02	0.727586	0.538561	0.393844	0.410408

表 10: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、PER 値）

	読み	読み、品詞	読み、正規	100%
set01	0.707746	0.490814	0.384924	0.387097
set02	0.698591	0.482495	0.356670	0.370589

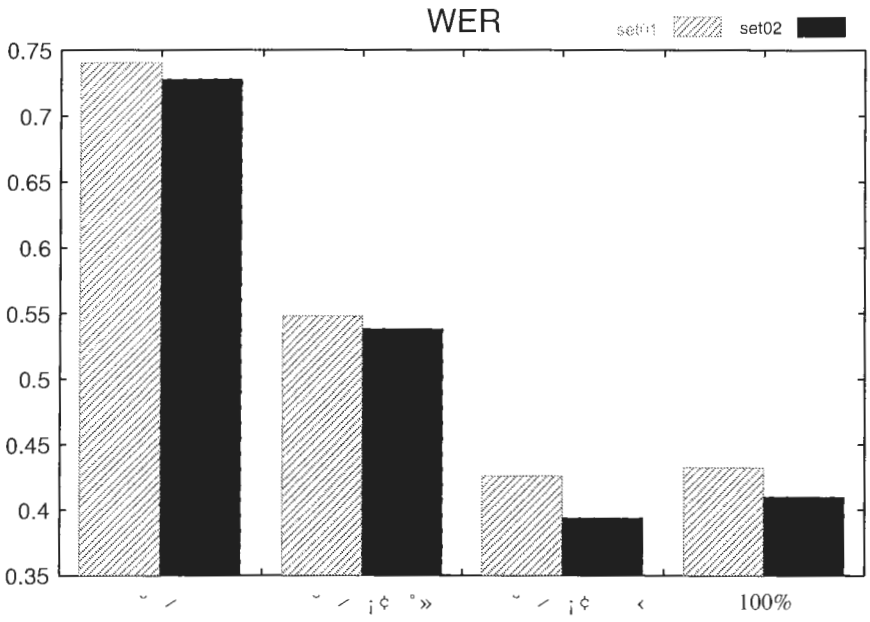

図 9: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、WER 値）

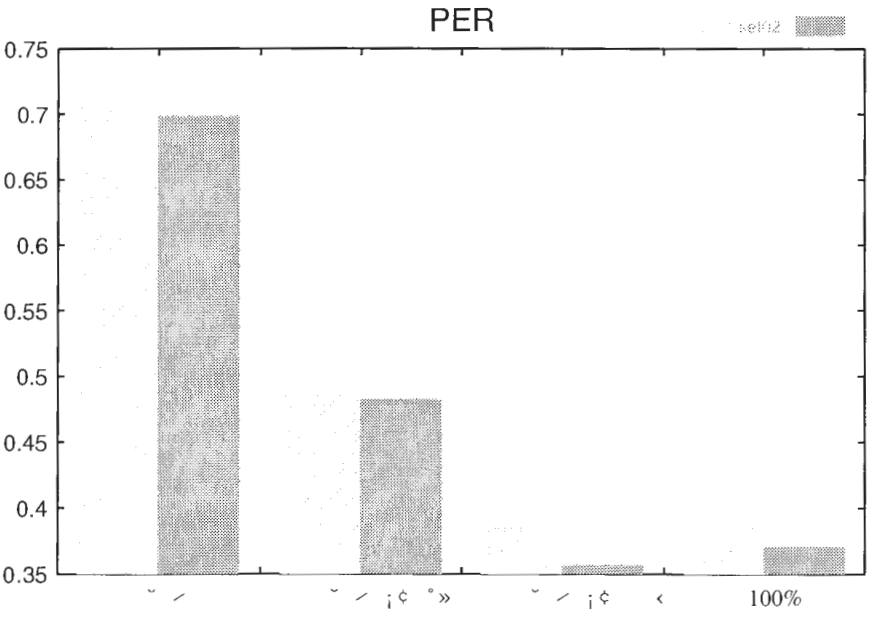

図 10: 実験結果（学習コーパスの単語属性の詳細度と翻訳精度の関係、PER 値）

4.1.2 考察

set01、02 の傾向は 02 の方がどの指標においても良い精度で翻訳されている。各コーパスの傾向を見てみると、まず「読み」だけのコーパスではほとんど翻訳できていないことがわかる。翻訳の出力を見てみると、何カ所かでは正しい単語が見られるが、文になっていないものが大半であった。「読み」の情報だけでは翻訳モデルの学習が十分に出来なかったのではないかと考えられる。

次に、「読み、品詞」のコーパスでは、「読み」だけのコーパスと比べて翻訳精度は向上してはいるが、100%のコーパスには及ばない。また、このコーパスで翻訳を行った場合、多義解消のタスクを暗に行ったことになるが、表 11（ここでは語＝形態素としている）に学習コーパスに含まれる同音語の数について見てみると、今回使用したコーパスには多義性のある語が多く含まれているわけではない。したがって、「読み、品詞」のコーパス学習から多義解消がうまく処理されたかどうかは判断出来ない。

表 11: 学習コーパス（100%）に含まれる同音語の数

	1	2	3	4	5～	合計
同音語の数	15339	893	184	80	54	16550

「読み、正規」のコーパスでは、100%コーパスを若干上回る評価となっている。これは、100%コーパスには、多くの情報が付与されているために、曖昧性が高くなっている可能性がある。元のコーパスではより詳細な情報があるために選択肢が増えてしまい、精度が下がったのかもしれない。また、「読み、品詞」のコーパスより、「読み、正規」のコーパスが精度が上回った。この結果だけを見ると、「品詞」情報が翻訳を行う際に役立っていないように思われるが、「正規」の情報には「品詞」情報は含まれていると考えられるため、一概に「品詞」情報が翻訳に必要のないものとは言い切れない。

5 まとめ

今回の調査では、「学習コーパス量と翻訳性能の関係」「学習コーパスにおける単語属性の詳細度と翻訳性能の関係」という 2 つの実験を行った。

「学習コーパス量と翻訳性能の関係」の実験では、学習コーパス量を変化させ翻訳を行い、それを評価することで、今回使用した SAT （Statistical ATR Translator）と学習コーパスである BTEC（旅行会話基本表現集）の学習量と翻訳精度の関係を明らかにした。また、言語モデルを固定し翻訳モデルの量だけを変化させたコーパスで翻訳を行い、翻訳精度への影響を調査した。今回の調査では、特に 60%コーパスにおいて著しく評価が下がる結果となった。これは、中途半端な学習量による翻訳性能の悪化、翻訳ソフトのパラメータとの相性の悪さ、などの理由が考えられる。

「学習コーパスにおける単語属性の詳細度と翻訳性能の関係」の実験では、音声翻訳における音声認識部と機械翻訳部の最適化するための調査として、入力の各単語に付与されている情報を削除し、学習・翻訳を行うことでどのような影響があるか調査した。「読み」情報のみでは全く翻訳を行うことが出来ず、「読み、正規」のコーパスでは 100%のオリジナルのコーパスより若干良い評価が得られた。これは、100%コーパスには多くの情報が付与されているために、訳語決定において曖昧性が生じてしまうためであると考えられる。また、「読み、品詞」のコーパスでの翻訳精度はそれなりに良いものであったが、今回使用したコーパスには多義性のある語が多く含まれておらず、機械翻訳部で多義解消のタスクを行えたかどうかはわからない、という結果となった。

参考文献

[1] 北 研二：確率的言語モデル、東京大学出版会 (1999)

[2] 長尾 真 編：岩波講座ソフトウェア科学 15 自然言語処理、岩波書店 (1996)

[3] KIshore Papineni,Salim Roukos, Todd Ward, and WeiJing Zhu: BLEU: a method for automatic evaluation of machine translation, in Proceedings of the 40th Annual Meeting of the Association for Comuputational Linguistics(ACL), pages 311-318 (2002)

[4] Geroge Doddington: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in Proceedings of the HLT Conference, San Diego, California (2002)

[5] 今村 賢治、隅田 英一郎、松本 裕治：機械翻訳自動評価指標の比較、第 10 回言語処理学会年次大会, pages 452-455 (2004)