TR-IT-0298

N-GRAM を用いた翻訳候補の選別

Study on Selecting and Sorting Candidates in
Translation Using N-GRAM

垣　　智
Satoshi Kaki

1999 年 3 月

内容概要

翻訳システムが出力した翻訳文の質の評価は、システムの性能向上や翻訳結果の後処理などを行う上で非常に重要である。現状では、このような評価は原言語と目的言語を一つ一つ比較しながら人間が行っており、かなりの時間と人手がかかっている。厳密ではなくてもある程度の精度で翻訳文の質を機械的にふるい分けることができれば、様々場面での応用が期待できる。そこで、本研究は N-GRAM を用いてこのような翻訳文の機械的判別精度について検討を行った。

ＡＴＲ音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

目　次

はじめに

翻訳システムが出力した翻訳文の質の評価は、システムの性能向上や翻訳結果の後処理などを行う上で非常に重要である。現状では、このような評価は原言語と目的言語を一つ一つ比較しながら人間が行っており、かなりの時間と人手がかかっている。厳密ではなくてもある程度の精度で翻訳文の質を機械的にふるい分けることができれば、様々場面での応用が期待できる。そこで、本研究は主に N-GRAM を用いてこのような翻訳文の機械的判別精度について検討を行った。検討には、特徴が異なる次の3種類のデータを用いた。

意味距離尤度同一値候補
翻訳システム訓練前後の翻訳文
科学技術文献抄録翻訳文

第１章　意味距離尤度同一値候補

1.1　概要

ＴＤＭＴ翻訳システムにおいて、翻訳処理の最終出力段階で意味距離尤度が同じ複数候補の生成がしばしば発生する。現システムでは、それら複数候補から最終翻訳結果への選択はまったくランダムにおこなっている。本実験では、各種 N-GRAM を用いて、このような意味距離尤度が同じ複数候補からもっともらしいものを選択する精度を求めた。

1.2　検証データ

検証データは、意味距離が同じ複数の訳文候補の中から目的言語に堪能な人間が、正解文と最良文を選出したものである。正解文は複数選択可能である。一方、最良文は正解文の中でもっとも良いものを一つ選択している。場合によっては、正解文、最良文とも該当するものがないものある（図 1-1 参照)。検証データの特徴を表 1-1 にまとめた。

```
"はい京都観光ホテルです"
090101 "Yes , I'm the Kyoto Kanko Hotel ."
090102 "Yes , this is the Kyoto Kanko Hotel ."
090103 "Yes , it's the Kyoto Kanko Hotel ."
090104 "Hello , I'm the Kyoto Kanko Hotel ."
090105 "Hello , this is the Kyoto Kanko Hotel ."
090106 "Hello , it's the Kyoto Kanko Hotel ."
正解文番号: 090102 090105
最良文番号: 090105
```

図 1-1　ＪＥ翻訳結果の人間評価シート例

判別精度は、検証データに対して相対評価予測と絶対評価予測を行った。相対評価予測は、任意の原言語に属する目的言語候補の中から最良と予測される候補を選出し、予測候補が人間評価と一致するか否かを検証したものである。一方、絶対評価予測は、原言語にかかわらず、すべての目的言語候補を対象に人間評価の判別結果をどれだけ予測できるかを検証したものである。

相対評価予測に関わる検証データの特徴は表の No.１～７が関係する。絶対評価予測については No.８～１２の項目である。

表 1-1　　検証データの特徴

No	項目	ＪＥ	ＥＪ			
			close	open	other	all
1	対象データ数（原言語）	99	188	55	54	297
2	正解文が存在したデータ数（原言語）	69	53	18	14	85
3	最良文が存在したデータ数（原言語）	47	88	22	24	134
4	正解文が存在するデータにおける平均候補文数	3.62	2.81	2.89	2.79	2.82
5	最良文が存在するデータにおける平均候補文数	3.38	2.67	2.5	2.46	2.60
6	候補文をランダムに一つ選択してそれが正解文である確率	41.51	48.68	47.04	44.64	47.67
7	候補文をランダムに一つ選択してそれが最良文である確率	29.56	37.45	40.00	40.68	38.40
8	全候補数（目的言語）	366	476	140	134	750
9	正解文として選出され候補文数（目的言語）	91	371	78	88	537
10	正解文として選出されなかった候補文数（目的言語）	275	105	62	46	213
11	最良文として選出され候補文数（目的言語）	47	88	22	24	134
12	最良文として選出されなかった候補文数（目的言語）	319	388	118	110	616

1.3 N-GRAMによる検討結果

文字 N-GRAM を用いた手法を検討した。この方法は翻訳文に含まれる連鎖単位の平均 N-GRAM 確率を用いて文の良否を数量化するものである。N-GRAM の連鎖単位は、文字と単語・品詞の混合 N-GRAM の二種類で実験を行った。用いたコーパスは下表のとおりである。ここで、ＪＥ、ＥＪは翻訳方向を示しており、ＪＥは日本語を英語に翻訳、ＥＪは英語を日本語に翻訳するものである。また、表中、ＪＥ、ＥＪに続く () 内の「文字」「混合」は、前者が文字 N-GRAM、後者が単語・品詞混合 N-GRAM であることを示している。

英語文字 N-GRAM の場合、文字種類はアルファベット（大文字小文字は区別している)、ピリオド、コンマなどである。また、単語間のスペースも一文字として扱っている。

日本語文字 N-GRAM は、文の形態素解析（ＪＵＭＡＮ）を行った後、数詞、固有名詞、人名などを記号で置き換えた抽象化を施している。

英語混合 N-GRAM は Black Tagger によって UPEN 体系のタグ付けを行い、以下の品詞単語は品詞扱いとし、それ以外は単語扱いとした。

VB, VBD, VBG, VBN, VBP, VBZ, NN, NNP, NNS, RB, JJ

日本語混合 N-GRAM は、ＪＵＭＡＮによる形態素解析結果に基づいて、助動詞、助詞は単語扱い、その外の単語は品詞扱いとした。

表 1-2　コーパス諸元

項目	ＪＥ（文字）	ＪＥ（混合）	ＥＪ（文字）	ＥＪ（混合）
文章数	10022	10022	14110	14110
延単位数	649797	146582	404110	222184
異なり単位数	60	517	1248	232

1.3.1 絶対評価予測

翻訳候補文のうち、人間が選出した正解文・最良文とそれ以外の候補文を原言語文に関わらずに判別する精度を求めた。N-GRAM 次元は２もしくは３から１０連鎖までを適用してみた。結果は下表のとおりで、数値は判別に成功した件数の全件数に対する割合（%）で示してある（ランダムに選んだ場合は５０％になる)。

表 1-3-1　ＪＥ（文字）の場合

N	3	4	5	6	7	8	9	１０
正解文	59.29	65.85	65.85	68.03	66.12	66.12	66.67	67.49
最良文	61.48	68.58	64.75	68.58	71.04	70.22	70.49	71.31

表 1-3-2　ＪＥ（混合）

N	2	3	4	5	6	7	8	9	１０
正解文	49.18	54.64	51.09	58.20	59.84	62.02	62.84	59.56	58.74
最良文	45.08	57.92	58.20	66.12	68.31	71.86	67.49	66.12	65.85

表 1-3-3　ＥＪ（文字）の場合

	N	2	3	4	5	6	7	8
close	正解文	74.14	62.39	59.66	57.14	57.98	59.87	60.08
	最良文	47.90	50.63	48.95	50.00	48.53	75.21	76.89
open	正解文	57.86	59.29	64.29	65.00	65.71	65.00	65.00
	最良文	52.14	54.29	57.14	58.57	60.71	57.86	57.86
other	正解文	50.00	55.22	53.73	53.73	52.24	56.72	47.76
	最良文	66.42	52.99	73.88	73.88	82.09	75.37	77.61
all	正解文	61.87	54.53	54.27	52.13	51.60	53.33	51.60
	最良文	35.47	27.60	22.80	22.13	29.20	55.20	56.53

表 1-3-4　　ＥＪ.close（混合）

N	2	3	4	5	6	7	8	9	10
正解文	53.99	52.10	53.99	55.67	55.46	51.26	52.73	55.04	52.10
最良文	18.49	27.52	23.95	35.71	35.50	40.34	40.55	39.71	35.50

●文字 N-GRAM は、人間選出基準とある程度相関がありそうである。一方、ＥＪの混合 N-GRAM にはほとんど相関はみられない。

●ＥＪ（文字）の場合をみると、close, open, other などを個別で判別する場合と合わせて判別する場合で判別率が大きく異なっている（特に最良文）。絶対評価予測では最良あるいは最悪な翻訳文によって判別基準位置が大きく影響を受ける傾向があるためと思われる。

1.3.2　相対評価予測

同一原言語内の候補文の中で、N-GRAM 確率のもっと良いものが、人間が選出した候補文と一致する割合を計測した。結果を下表に示す。表中の RAND 欄がランダムに選出した場合の確率である。

表 1-4-1　　ＪＥ（文字）の場合

N	3	4	5	6	7	8	9	10	RAND
正解文	48.44	64.06	46.88	54.69	59.38	53.13	54.69	60.94	41.51
最良文	40.43	48.94	44.68	51.06	57.45	51.06	55.32	57.45	29.56

表 1-4-2　　ＪＥ（混合）の場合

N	2	3	4	5	6	7	8	9	10	RAND
正解文	42.19	50.00	51.56	62.50	57.81	54.69	53.12	48.44	48.44	41.51
最良文	42.55	42.55	51.06	61.70	51.06	53.19	48.94	42.55	42.55	29.56

表 1-4-3　　ＥＪ（文字）の場合

	N	2	3	4	5	6	7	8	RAND
close	正解文	62.26	62.26	69.81	73.58	73.58	66.04	69.81	48.68
	最良文	67.05	62.50	73.86	73.86	69.32	70.45	64.77	37.45
open	正解文	55.56	61.11	55.56	61.11	66.67	61.11	61.11	47.04
	最良文	54.55	54.55	54.55	63.64	63.64	59.09	59.09	40.00
other	正解文	50.00	64.29	71.43	64.29	64.29	64.29	57.14	44.64
	最良文	66.67	66.67	70.83	62.50	70.83	70.83	62.50	40.68
all	正解文	58.82	62.35	67.06	69.41	70.59	64.71	65.88	47.67
	最良文	64.93	61.94	70.15	70.15	68.66	68.66	63.43	38.40

表 1-4-4　　ＥＪ.close（混合）

N	2	3	4	5	6	7	8	9	10	RAND
正解文	54.72	52.83	50.94	54.72	67.92	60.38	58.49	58.49	60.38	48.68
最良文	53.41	46.59	44.32	46.59	65.91	60.23	63.64	63.64	62.50	37.45

●いずれの場合もランダムに選ぶ割合（RAND 欄の値）を超えている。

●ＥＪの場合、文字と混合 N-GRAM では、文字 N-GRAM の値が常に高い値を示している。

●英語と日本語で比較すると日本語の方が高い値を示している。ただ、ランダム値との差

で比較するとあまり変わらない。

●ＥＪ（文字）の場合、N-GRAM 次元が４から６付近に最高値が集まっている。

図 1-2-1 から 1-2-3 に解析例を示す。どの例も N-GRAM の次元が高くなるにしたがって未知連鎖が増えている。したがって、次元がある程度高くなると、未知連鎖の多発によって、ほとんどの場合文頭あるいは文末の表現でしか N-GRAM 平均に違いは現れていない。一方、ＪＥの場合は、異なり文字数が少ないことから未知連鎖の発生はＥＪに比較してかなり少ない。

図 1-1-1 の例を見ると、各次元での一位から二位候補が最良文あるいは正解文に一致していることがわかる。

一方、図 1-2-2 では、次元が変わるに従がって一位候補がくるくる入れ替わっている。これは、次元が極端に低い場合は極々近傍しか見ない、一方、次元が高い場合には上述したように文頭、文末表現と限られた定型表現しか見ないということが原因と思われ、相対評価予測には適度な N-GRAM 次元が必要である。

図 1-2-1　　ＥＪ（文字）解析例

【順位の遷移状況】

SNo.	正解文	最良文	N						
			2	3	4	5	6	7	8
1			6	5	8	7	7	7	7
2	○		4	2	2	1	3	3	3
3			5	6	6	5	8	8	8
4			8	8	5	3	4	4	4
5			3	3	7	8	5	5	5
6	○	○	1	1	1	2	1	1	1
7			2	4	4	6	6	6	6
8	○		7	7	3	4	2	2	2

【N-GRAM 値】

N	SNo							
	1	2	3	4	5	6	7	8
2	-1.802705	-1.792440	-1.802072	-1.867140	-1.760657	-1.750392	-1.760024	-1.825092
3	-1.977041	-1.887658	-1.977041	-2.134175	-1.917680	-1.828297	-1.917680	-2.074815
4	-2.915020	-2.537985	-2.782100	-2.781502	-2.830256	-2.453221	-2.697335	-2.696737
5	-3.701071	-3.308917	-3.570353	-3.439635	-3.794438	-3.402284	-3.663720	-3.533002
6	-4.223679	-3.978766	-4.223679	-3.978766	-4.189471	-3.944558	-4.189471	-3.944558
7	-4.453144	-4.216235	-4.453144	-4.216235	-4.299699	-4.062790	-4.299699	-4.062790
8	-4.445384	-4.211314	-4.445384	-4.211314	-4.298780	-4.064711	-4.298780	-4.064711

N-GRAM の最大、最小区間を５分割して記号で表示した。
「＊」最も良い
「＋」良い
「　」普通
「－」悪い
「＝」最も悪い
「Ｘ」未知連鎖

(S1)TXT->あなたが名前と到着することができる番号を教えるなら、はい、予約を完了することができます。
(02)NGV->　　X＊＊―＊＋＋＋＊＋＋＋＋X＊＊＋＊＋　＋＊　＊＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->　＊XXX＊X＊＊＊＊＊＊＊＊＊XX＊＊＊X　＊＊　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->　＊XXXXXXX＊X＊＊＊＊＊XXX＊＊XX＊＊XX＊　＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->　＊XXXXXXXXXX＊＊＊＊XXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->　＊XXXXXXXXXXX＊＊＊XXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->　＊XXXXXXXXXXXX＊＊XXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＋　＊XXXXXXXXXXXXX＋XXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S2)TXT->あなたが名前と連絡することができる番号を教えるなら、はい、予約を完了することができます。
(02)NGV->　　X＊＊　＊＋＋＋＊＋＋＋X＊＊＋＊＋　＋＊　＊＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->　＊XXX＊＋＊　＊＊＊＊＊＊＊XX＊＊＊X　＊＊　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->　＊XXXX＋＊X＊＊＊＊＊＊＊XXX＊＊XX＊＊XX＊　＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->　＊XXXXX＊XX＊＊＊＊＊＊XXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->　＊XXXXXXXXX＊＊＊＊XXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->　＊XXXXXXXXXXX＊＊＊XXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＋　＊XXXXXXXXXXXX＊＊＋XXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S3)TXT->あなたが到着することができる名前と番号を教えるなら、はい、予約を完了することができます。
(02)NGV->　　―＊＋＋＋＊＋＋＋X＊＊X＊＊＊＊＋　＋＊　＊＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->　＊XX＊＊＊＊＊＊＊＊XX＊XX＊＊＊X　＊＊　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->　＊XXX＊＊X＊＊＊＊XXXXXX＊＊XX＊＊XX＊　＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->　＊XXXX＊XX＊＊＊＊XXXXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->　＊XXXXXXXX＊＊＊XXXXXXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->　＊XXXXXXXXX＊＊XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＋　＊XXXXXXXXXX＋XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S4)TXT->あなたが連絡することができる名前と番号を教えるなら、はい、予約を完了することができます。
(02)NGV->　　X＊＋＋＊＋＋＋X＊＊X＊＊＊＊＋　＋＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->　＊XXX　＊＊＊＊＊＊＊XX＊XX＊＊＊X　＊＊　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->　＊XXXX＊＊＊＊＊＊XXXXXX＊＊XX＊＊XX＊　＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->　＊XXXXX＊＊＊＊＊XXXXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->　＊XXXXXX＊＊＊＊XXXXXXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->　＊XXXXXXX＊＊＊XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＋　＊XXXXXXXX＊＊＋XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S5)TXT->はい、あなたが名前と到着することができる番号を教えるなら、予約を完了することができます。
(02)NGV->＋＊＊＋　　X＊＊―＊＋＋＋＊＋＋＋X＊＊＋＊＋　＋＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->＋＊＊＋＋＊XXX＊X＊＊＊＊＊＊＊＊＊XX＊＊＊X　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->＋＊＊＋X＊XXXXXX＊X＊＊＊＊＊XXX＊＊XX＊＊XX＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->＋＊＊＋XXXXXXXXXXX＊＊＊XXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->＋＊＊＋XXXXXXXXXXXX＊＊＊XXXXXXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->＋＊＊＋XXXXXXXXXXXXX＊＊XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＊＊＊　XXXXXXXXXXXXXX＋XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S6)TXT->はい、あなたが名前と連絡することができる番号を教えるなら、予約を完了することができます。
(02)NGV->＋＊＊＋　　X＊＊　＊＋＋＋＊＋＋＋X＊＊＋＊＋　＋＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->＋＊＊＋＋＊XXX＊＋＊　＊＊＊＊＊＊＊XX＊＊＊X　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->＋＊＊＋X＊XXXX＋＊X＊＊＊＊＊＊＊XXX＊＊XX＊＊XX＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->＋＊＊＋XXXXXXX＊XX＊＊＊＊＊＊XXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->＋＊＊＋XXXXXXXXXXX＊＊＊＊XXXXXXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->＋＊＊＋XXXXXXXXXXXX＊＊＊XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＊＊＊　XXXXXXXXXXXXX＊＊＋XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S7)TXT->はい、あなたが到着することができる名前と番号を教えるなら、予約を完了することができます。
(02)NGV->＋＊＊＋　　―＊＋＋＋＊＋＋＋X＊＊X＊＊＊＊＋　＋＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->＋＊＊＋＋＊XX＊＊＊＊＊＊＊＊＊XX＊XX＊＊＊X　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->＋＊＊＋X＊XXX＊＊X＊＊＊＊＊XXXXXX＊＊XX＊＊XX＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->＋＊＊＋XXXXXX＊XX＊＊＊＊XXXXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->＋＊＊＋XXXXXXXXXX＊＊＊XXXXXXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->＋＊＊＋XXXXXXXXXXX＊＊XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊
(08)NGV->＊＊＊　XXXXXXXXXXXX＋XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊
(S8)TXT->はい、あなたが連絡することができる名前と番号を教えるなら、予約を完了することができます。
(02)NGV->＋＊＊＋　　X＊＋＋＋＊＋＋＋X＊＊X＊＊＊＊＋　＋＊　＊＊XXX＋＋＊＋＋＋＊＊＊＊
(03)NGV->＋＊＊＋＋＊XXX　＊＊＊＊＊＊＊XX＊XX＊＊＊X　＊＊　＊＊XXXX＊＊＊＊＊＊＊＊＊
(04)NGV->＋＊＊＋X＊XXXX＊＊＊＊＊＊＊XXXXXX＊＊XX＊＊XX＊＊XXXXX＊＊＊＊＊＊＊＊
(05)NGV->＋＊＊＋XXXXXXX＊＊＊＊＊＊XXXXXX＊XXX＊XXXX＊＊XXXXXX＊＊＊＊＊＊＊
(06)NGV->＋＊＊＋XXXXXXXX＊＊＊＊＊XXXXXXXXXXXXXXXXX＊XXXXXXX＊＊＊＊＊＊
(07)NGV->＋＊＊＋XXXXXXXXX＊＊＊＊XXXXXXXXXXXXXXXXXXXXXXXXX＊＊＊＊＊

(08)NGV->＊＊＊　ＸＸＸＸＸＸＸＸＸＸ＊＊＋ＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸ＊＊＊＊

図 1-2-2　　ＥＪ（文字）解析例

【順位の遷移状況】

SNo.	正解文	最良文	N						
			2	3	4	5	6	7	8
1	○	○	2	1	1	1	1	1	2
2			1	2	2	2	2	2	1

【N-GRAM 値】

N	SNo	
	1	2
2	-1.687127	-1.685325
3	-2.112254	-2.132694
4	-3.185327	-3.242004
5	-3.994040	-4.235968
6	-4.266387	-4.504285
7	-4.478810	-4.535203
8	-4.721658	-4.611937

(S1)TXT->では、私達@N 人に今日のスペシャルを注文する方が良いと思います。
(02)NGV->　＋＊—ＸＸ＋＋—＋＊　＋＊＊＊＋　＊＋＋＋＊　＊＋＊＊＊＊＊＊
(03)NGV->　＊＊ＸＸＸＸ＋ＸＸ＋＋＊＊＊＊ＸＸ＊＋＊＋＊＋＊＋＊＊＊＊＊＊
(04)NGV->　＊＊ＸＸＸＸＸＸＸＸＸ＊＊＊＊ＸＸＸＸ＊Ｘ＊Ｘ＊Ｘ＊＊＊＊＊＊
(05)NGV->　＊＊ＸＸＸＸＸＸＸＸＸＸ＊＊＊ＸＸＸＸＸＸＸＸＸＸＸ＊＊＊＊＊
(06)NGV->　＊＊ＸＸＸＸＸＸＸＸＸＸＸ＊＊ＸＸＸＸＸＸＸＸＸＸＸＸ＊＊＊＊
(07)NGV->＋＊＊ＸＸＸＸＸＸＸＸＸＸＸＸ＊ＸＸＸＸＸＸＸＸＸＸＸＸＸ＊＊＊
(08)NGV->＋＊＊ＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸ＊
(S2)TXT->では、今日の私達@N 人にスペシャルを注文する方が良いと思います。
(02)NGV->　＋＊　＋＊—ＸＸ＋＋　＋＊＊＊＋　＊＋＋＋＊　＊＋＊＊＊＊＊＊
(03)NGV->　＊＊＋＋＋ＸＸＸＸ＋ＸＸ＊＊＊ＸＸ＊＋＊＋＊＋＊＋＊＊＊＊＊＊
(04)NGV->　＊＊　Ｘ＊ＸＸＸＸＸＸＸＸ＊＊ＸＸＸＸ＊Ｘ＊Ｘ＊Ｘ＊＊＊＊＊＊
(05)NGV->　＊＊＋ＸＸＸＸＸＸＸＸＸＸＸ＊ＸＸＸＸＸＸＸＸＸＸＸ＊＊＊＊＊
(06)NGV->　＊＊＋ＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸ＊＊＊＊
(07)NGV->＋＊＊＋ＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸ＊＊＊
(08)NGV->＋＊＊　ＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸＸ＊

図 1-2-3　　ＪＥ（文字）解析例

【順位の遷移状況】

Sno	正解文	最良文	N							
			3	4	5	6	7	8	9	10
1			3	4	4	4	4	3	3	3
2			1	3	3	2	2	2	1	1
3			4	2	2	3	3	4	4	4
4	○	○	2	1	1	1	1	1	2	2

(S1)TXT->I'd like to have a room with a view in front a sea, please and.
(03)NGV->+***+*******+**********+***+**+*+**+**+**++*+*+*+ *+*********+*
(04)NGV->+**********+**********+******+****+**+*****+*+**X*+******+**+*
(05)NGV->+**********+**********+****+**+***** *+**XX*******+**X*
(06)NGV->+**********+*******+**********+****+**X***** X+**XXX******+**XX
(07)NGV->+**********+******+***X+**XX**** XXX*XXXX*****+**XX
(08)NGV->+**********+***XX**XXX*** XXXXXXXXX****+**XX
(09)NGV->+**********+***XXX*XXXX**+XXXXXXXXXX***+**XX
(10)NGV->+**********+**********+******+***XXXXXXXXX*+XXXXXXXXXX*X+**XX
(S2)TXT->I'd like to have a room with a view in front of the sea, please and.
(03)NGV->+***+*******+**********+***+**+*+**+**+**++*+******+*+ *+*********+*
(04)NGV->+**********+**********+******+****+**+*****+******+**X*+******+**+*
(05)NGV->+**********+**********+****+**+**********+**XX*******+**X*
(06)NGV->+**********+*******+**********+****+**X**********+**XXX******+**XX
(07)NGV->+**********+******+***X+**XX**********+**XXXX*****+**XX
(08)NGV->+**********+***XX**XXX**********+**XXXXX****+**XX
(09)NGV->+**********+***XXX*XXXX*********+*XXXXXXX***+**XX
(10)NGV->+**********+**********+******+***XXXXXXXXX********+*XXXXXXX*X+**XX
(S3)TXT->I'd like to have a room with a view in front a sea and, please.
(03)NGV->+***+*******+**********+***+**+*+**+**+**++*+*+*+++**+*+*****+*
(04)NGV->+**********+**********+******+****+**+*****+*+** *** *********
(05)NGV->+**********+**********+****+**+***** *+** X** *+*******
(06)NGV->+**********+*******+**********+****+**X***** X+**XXX*+*X*******
(07)NGV->+**********+******+***X+**XX**** XXX*XXXX**XX******
(08)NGV->+**********+***XX**XXX*** XXXXXXXXX*XXX*****
(09)NGV->+**********+***XXX*XXXX**+XXXXXXXXXXXXXX****
(10)NGV->+**********+**********+******+***XXXXXXXXX*+XXXXXXXXXXXXXX***
(S4)TXT->I'd like to have a room with a view in front of the sea and, please.
(03)NGV->+***+*******+**********+***+**+*+**+**+**++*+******+*+++**+*+*****+*
(04)NGV->+**********+**********+******+****+**+*****+******+** *** *********
(05)NGV->+**********+**********+****+**+**********+** X** *+*******
(06)NGV->+**********+*******+**********+****+**X**********+**+XX*+*X*******
(07)NGV->+**********+******+***X+**XX**********+**+XXX**XX******
(08)NGV->+**********+***XX**XXX**********+**+XXXX*XXX*****
(09)NGV->+**********+***XXX*XXXX*********+*X+XXXXXXXX****
(10)NGV->+**********+**********+******+***XXXXXXXX********+*XXXXXXXXXXX***

1.3.3　相対評価予測におけるコーパス量効果

相対評価予測におけるコーパス量の影響を調べた。結果を表 1-5-1 から 1-5-3 に示す。

ＪＥ（文字）では N-GRAM 次元が高く、コーパス量が増えるほどに判別精度が上昇している。一方、ＥＪ（文字）では、正解文の予測についてはＪＥ（文字）と同様の傾向を示すが、最良文ではコーパス量に関わらず、次元が一定の位置に判別精度の高い領域が集まっている。ＥＪ（混合）の場合、正解文予測には一定の傾向は見られないが、最良文予測はＪＥ（文字）に近い傾向を示している。

これらの結果は異なり単位数の違いによるものと考えられる。異なり単位数が多い場合は、分野をしぼったコーパスで対処するのが実用的であろう。

表 1-5-1(1)　　ＪＥ（文字）、正解文

文数	延文字数	異なり文字数	3	4	5	6	7	8	9	10
3691	231585	60	43.75	54.69	51.56	48.44	50.00	46.88	50.00	56.25
4101	257841	61	43.75	56.25	51.56	53.13	51.56	48.44	53.13	54.69
4613	290093	61	43.75	53.13	51.56	53.13	51.56	48.44	54.69	56.25
5272	330573	61	43.75	51.56	53.13	57.81	56.25	53.13	56.25	56.25
6151	385326	61	45.31	54.69	53.13	60.94	59.38	56.25	57.81	59.38
7381	459550	61	43.75	56.25	50.00	64.06	62.50	60.94	59.38	59.38
9226	573554	61	46.88	57.81	51.56	57.81	64.06	53.13	56.25	59.38
12302	764074	62	45.31	54.69	45.31	50.00	60.94	62.50	62.50	65.63
18452	1141087	62	45.31	56.25	48.44	54.69	62.50	67.19	64.06	67.19

表 1-5-1(2)　　ＪＥ（文字）、最良文

3	4	5	6	7	8	9	10
38.30	46.81	51.06	46.81	51.06	48.94	48.94	55.32
38.30	44.68	51.06	48.94	51.06	48.94	51.06	51.06
38.30	42.55	46.81	44.68	51.06	48.94	53.19	53.19
38.30	40.43	48.94	48.94	53.19	51.06	51.06	51.06
38.30	42.55	48.94	55.32	55.32	53.19	55.32	55.32
38.30	44.68	48.94	57.45	55.32	55.32	55.32	57.45
38.30	44.68	48.94	51.06	61.70	51.06	55.32	59.57
38.30	44.68	44.68	46.81	57.45	57.45	59.57	63.83
40.43	42.55	44.68	53.19	59.57	65.96	61.70	63.83

表 1-5-2(1)　　E J .close（文字)、正解文

文数	延文字数	異なり文字数	2	3	4	5	6	7	8
4077	114170	1104	52.83	62.26	62.26	56.60	54.72	60.38	58.49
4530	126850	1128	52.83	58.49	60.38	54.72	54.72	60.38	58.49
5097	142880	1158	54.72	56.60	62.26	60.38	58.49	64.15	62.26
5825	163455	1195	54.72	56.60	60.38	58.49	60.38	62.26	62.26
6795	190298	1244	62.26	54.72	60.38	62.26	60.38	56.60	64.15
8154	228652	1280	60.38	58.49	62.26	62.26	64.15	58.49	66.04
10193	286767	1335	62.26	56.60	62.26	56.60	64.15	62.26	67.92
13590	382041	1409	58.49	54.72	56.60	54.72	62.26	64.15	69.81
20385	573857	1481	64.15	56.60	66.04	66.04	67.92	62.26	69.81

表 1-5-2(2)　　E J .close（文字)、最良文

2	3	4	5	6	7	8
60.23	69.32	71.59	70.45	62.50	56.82	55.68
57.95	67.05	70.45	69.32	62.50	56.82	54.55
57.95	65.91	70.45	72.73	64.77	59.09	56.82
59.09	69.32	71.59	71.59	65.91	56.82	56.82
62.50	68.18	71.59	73.86	65.91	56.82	56.82
61.36	70.45	72.73	73.86	68.18	61.36	57.95
63.64	68.18	72.73	68.18	65.91	63.64	59.09
63.64	67.05	69.32	65.91	65.91	64.77	61.36
63.64	65.91	72.73	70.45	69.32	64.77	62.50

表 1-5-3(1)　　E J .close（混合)、正解文

文章数	延形態素	異形態素	2	3	4	5	6	7	8	9	1 0
4077	204360	194	62.26	50.94	56.60	64.15	67.92	67.92	64.15	60.38	60.38
4530	228822	200	60.38	50.94	54.72	58.49	62.26	64.15	62.26	58.49	60.38
5097	259670	202	64.15	50.94	58.49	62.26	60.38	64.15	62.26	58.49	60.38
5825	300438	206	62.26	52.83	56.60	60.38	58.49	62.26	62.26	58.49	60.38
6795	354758	220	64.15	52.83	56.60	62.26	56.60	58.49	58.49	58.49	58.49
8154	438646	226	62.26	54.72	58.49	64.15	58.49	58.49	62.26	62.26	62.26
10193	569142	233	62.26	58.49	58.49	58.49	52.83	56.60	64.15	64.15	66.04
13590	786550	242	58.49	64.15	58.49	62.26	56.60	60.38	64.15	64.15	66.04

表 1-5-3(2)　　E J .close（混合)、最良文

2	3	4	5	6	7	8	9	1 0
56.82	47.73	51.14	61.36	63.64	64.77	64.77	60.23	60.23
56.82	46.59	52.27	61.36	61.36	63.64	64.77	61.36	61.36
57.95	47.73	54.55	62.50	60.23	63.64	64.77	61.36	61.36
55.68	48.86	54.55	59.09	59.09	62.50	63.64	61.36	61.36
57.95	48.86	52.27	59.09	56.82	59.09	60.23	60.23	59.09
57.95	51.14	48.86	56.82	59.09	60.23	64.77	64.77	63.64
55.68	53.41	48.86	50.00	53.41	56.82	64.77	64.77	64.77
53.41	59.09	54.55	54.55	55.68	62.50	67.05	64.77	64.77

第２章　訓練前後の翻訳データ

2.1　概要

ＴＤＭＴ翻訳システムにおいては、随時、性能向上を計るため訓練文を用いて翻訳ルールの追加改善を行っている。一般的に訓練後の翻訳結果が訓練前のものより質の良い訳文と見做すことができる。そこで、同じ原言語の原文に対して訓練前と訓練後で翻訳結果が異なるものを集め、この両者が機械的に選別できるかどうか実験を行った。

2.2　検証データ

検証に用いたＥＪデータの件数は２２６２件で、それぞれに訓練前と訓練後の日本語訳がある（図 2-1 参照）。

判別精度は、同じ原文内での相対評価予測と、原文に依存しない絶対評価予測を行って求めた。

図 2-1　　訓練前後データ例 *(上段が訓練前、下段が訓練後の訳文)*

```
Yes but we'll need at least a day to make arrangements
はいしかし必要少なくとも手配をするために日です
はいしかし手配をするために最低一日が必要です

And the fee for the party is forty five dollars per adult fifteen dollars per child including tax
それからパーティーの料金は一大人四十五ドル一子供十五ドルを税含みます
それからパーティーの料金は税を含めて大人四十五ドル子供十五ドルです

How long does it take round trip
どのくらい長くそれは往復を撮りますか
どのくらい長く往復にかかりますか
```

2.3　N-GRAM による検討結果

用いた N-GRAM は文字 3-GRAM で、使用したコーパスは下表の 4 種類である。コーパス及び翻訳文の日本語は形態素解析（ＪＵＭＡＮ）を行った後に、数詞、固有名詞などを記号で置き換えている。

表 2-1　　コーパス諸元

種類	標準	2 倍	標準＋新聞	標準＋新聞 2
文章数	20199	40895	38897	119651
延文字数	519901	956871	1374836	5353538
異なり文字数	1286	1442	2988	4424

予測の説明変数として次の 6 種類を検討した。

表 2-2　　説明変数

説明変数 No	説明変数	内容
1	文字長	翻訳文の文字長さ
2	誤り個数	文字毎の 3-GRAM 確率値において、一定の閾値以下の確率値を示す部分を誤りとみなし、その個数をカウントしたもの。
3	誤り文字長さ	各誤りに含まれる文字長さ合計。
4	文の N-GRAM 確率	一文における各文字の N-GRAM 確率値の平均値。
5	一文字あたりの誤り個数	誤り個数　／　文字長
6	一文字あたりの誤り文字長	誤り文字長　／　文字長

2.3.1 相対評価予測

訓練後の翻訳文が訓練前より優れているという前提で、訓練前後の翻訳文を機械的に相対的評価予測できるかどうかを調べた。結果を下表に示す。表中の判別率はまったくランダムに選択した場合、５０％となる。

説明変数「文の 3-GRAM 確率」の判別率が最も良く、約６６％の値を示した。それ以外の説明変数では５０％前後かあるいはそれ以下である。コーパスが二倍になると数ポイントではあるが判別率が上昇している。一方、新聞記事を加えてもほとんど変化が見られなかった。

表 2-3　相対評価予測の判別率

説明変数	標準	2 倍	標準＋新聞	標準＋新聞 2
誤り個数	36.0	38.1	35.9	37.4
誤り文字長さ	45.9	47.9	45.6	44.9
文字数当たりの誤り個数	54.1	54.7	53.9	52.2
文字数当たりの誤り文字長	57.0	57.9	56.9	54.4
文の 3-GRAM 確率（文字数で正規化）	65.9	66.7	66.9	66.3

相対評価予測における各説明変数の寄与度合いを求めてみると、「文の 3-GRAM 確率」がやはり一番寄与していることがわかる（表 2-4 参照)。「文の 3-GRAM 確率」と「誤り個数」の違いは、誤り個数（長）が一定以下の N-GRAM 確率部分のみを注目するのに対し、「文の 3-GRAM 確率」ではそれに加え N-GRAM 確率の高い部分（滑らかな文字列）をも評価しているためであろう。

表 2-4　相対評価予測における説明変数の寄与度合

No	説明変数	スコアレンジ	偏相関係数
1	文字長さ	0.003448038	0.021630985
2	誤り個数	0.001994502	0.005645530
3	誤り文字長さ	0.000648047	0.002227689
4	文あたりの確率	0.032221578	0.231041380
5	文字あたりの誤り個数	0.005675676	0.021513535
6	文字あたりの誤り文字長さ	0.004455568	0.017736820

（スコアレンジ、偏相関係数とも値が大きいほど寄与度合いが高い）

次に文字の連鎖方向と未知連鎖確率の設定を変化させた実験を行ったが、両者とも判別に影響を与えるものではなかった（表 2-5 参照)。

表 2-5　連鎖方向と未知連鎖確率の設定

（コーパスは標準、説明変数は「文の 3-GRAM 確率」）

変数	率 1	率 2
前方向	66.0	66.2
後方向	66.1	66.9
前＋後	66.2	66.5

率 1 ：未知連鎖確率を最低確率－ 2
率 2 ：未知連鎖確率を最低確率

３．２　絶対評価予測

翻訳文がどの原文からの訳かを考慮せず、訓練前と訓練後だけを判別する絶対評価予測を行った結果を下表に示す。ほぼ、５０％強の値を示し、ランダム選別した場合の５０％と同程度の値である。

表 2-6-1　　絶対評価予測（一説明変数の判別率）

説明変数	1	2	3	4	5	6
判別率	50.80	52.10	52.34	55.35	55.00	54.71

表 2-6-2　　絶対評価予測（二説明変数の判別率）

	2	3	4	5	6
1	54.44	55.22	55.86	54.93	54.53
2		52.67	55.46	54.80	54.22
3			55.44	54.95	54.75
4				55.44	55.02
5					54.75

第３章　科学文献抄録翻訳データ

3.1　概要

ＴＤＭＴシステムの応用として、科学技術文献抄録の翻訳プロジェクトがある。文献抄録の英語から日本語への翻訳である。大量の翻訳結果を評価する必要があり、その品質評価として次の３段階がある。（ａ）正しく翻訳されており後編集が不要、（ｂ）後編集に役立つ、（ｃ）再翻訳が必要。また、ニーズとして、後編集に役立つ場合、編集補助となる情報を編集者に提供する。なども考えられる。

本実験では、前者の品質評価の３区分を機械的に選別可能かどうかを検討した。

3.2　検証データ

検証用データとして、下図にあるような抄録翻訳結果を人間が評価したデータを用いた。この評価では、複数の人間によって、原文と出力された翻訳文を比較しながら３段階の評価を加えている。

図 3-1　　付録　人間評価の例

E=Computer environment is assumed in a way that a personal computer is connected with a supercomputer through LAN.
ANS=コンピュータ環境はパソコンからＬＡＮを介しスーパーコンピュータに接続することを想定した。

J=b|b=b|b|b=コンピュータ環境はＬＡＮでスーパーコンピュータで接続されたパソコンがあるという方法の中で仮定されています。
J=b|b|a|a=-|a|b=コンピュータ環境はパソコンがＬＡＮを通してスーパーコンピュータと接続されるという方法の中で仮定されています。
J=b|b|b|b|b|b|b|b=b|b|b=コンピュータ環境は方法に仮定されている‥そのパソコンはＬＡＮを通してスーパーコンピュータと接続されます。
J=c|b=-|b|c=コンピュータ環境は方法に仮定されている‥そのパソコンは接続される‥スーパーコンピュータからＬＡＮです。

E=Desires of Fluid Power Engineers to the Simulation Tool and the Introduction of available Ones.
ANS=油空圧技術者が望むシミュレーションツールと各種ソフトウエアの紹介。

J=c|c=c|c|c=利用可能なもののシュミレーションへの油空圧技術者の望む、ツールと導入です。
J=b|c=-|b|c=利用可能なもののシュミレーションへの流体の力技術者の望む、ツールと導入です。
J=b|b|b|c|b|b|b|b|b=-|b|b=利用可能なもののシュミレーションツールの油空圧技術者の望むのと導入です。
J=b|b|b=b|b|b=利用可能なもののシュミレーションツールの油空圧技術者や導入の望みます。

その分布をまとめたものが次の表に示す。表中で全員一致、多数決（甘口）、（辛口）とは、複数の人間の評価値は完全に一致したものを全員一致とした。全員一致では評価が別れるものは解析対象から除外している。一方、多数決はもっとも数が多い評価値を採用するものであるが、多数決の評価値が複数或場合に、評価が良い評価値を採用したのが甘口、評価が悪い評価値を採用したのが辛口である。今回の解析では、全員一致データを対象とした。

表 3-1　　人間評価の頻度分布

評価ﾗﾝｸ	全員一致	多数決（甘口）	多数決（辛口）
a（編集不要）	4	13	7
b（編集必要）	83	177	130
c（再翻訳必要）	19	28	81
-（分類不能）	112	0	0
合計	218	218	218

3.3　N-GRAMによる検討結果

文字 N-GRAM を用いて機械的に評価値を判別できるか検討を行った。コーパス、及び翻訳文は形態素解析（ＪＵＭＡＮ）を行った後、数詞、固有名詞、人名などを記号で置き換えて抽象化を施している。また、用いたコーパスも分野によって区分して検討を行った（下表参照）。

表 3-2　　コーパス諸元

種類	文章数	延文字数	異なり文字数
1)標準	2495	94112	1596
2)情報工学分野	2790	117433	1631
3)医療分野	1784	72668	1292
1)+2)	5285	211545	2003
1)+3)	4279	166780	1875
1)+2)+3)	7069	284213	2176

評価方法は、同一原言語内での訳文候補間の比較（相対評価予測）と原言語に関わらずに翻訳文全体についての評価値の判別を行う絶対評価予測があるが、検証用データの人間評価値分布に非常に偏りがみられるため、今回は絶対評価予測のみを行った。

3.3.1　一変数での判別率

判別に用いた説明変数は次の６種類である。未知の並びは最低確率よりさらに－２した値、誤り検出はの閾値は最低確率－１とした。

表 3-3　　説明変数

説明変数 No	説明変数	内容
1	文字長	目的言語の文字長さ
2	誤り個数	文字毎の３-GRAM 確率値において、一定の閾値以下の確率値を示す部分を誤りとみなし、その個数をカウントしたもの。
3	誤り文字長さ	各誤りに含まれる文字長さ合計。
4	文の N-GRAM 確率	一文における各文字の N-GRAM 確率値の平均値。
5	一文字あたりの誤り個数	誤り個数　／　文字長
6	一文字あたりの誤り文字長	誤り文字長　／　文字長

表 3-4　　判別率

ｺｰﾊﾟｽ/説明変数	1	2	3	4	5	6
1	33.96	46.23	35.85	50.00	33.96	30.19
2	33.96	24.53	37.74	29.25	28.30	41.51
3	33.96	19.81	29.25	9.44	41.51	16.98
1＋2	33.96	40.57	29.25	51.89	39.62	30.19
1＋3	33.96	38.69	33.96	47.17	33.02	21.70
1＋2＋3	33.96	37.74	30.19	49.06	44.34	31.13

3.3.2 複数説明変数での判別率

表 3-5-1　判別率（コーパス：標準）

説明変数	2	3	4	5	6
1	52.83	50.94	50.94	60.38	53.77
2		54.72	57.55	56.60	52.83
3			51.89	50.00	51.89
4				33.96	35.85
5					40.57

表 3-5-2　判別率（コーパス：情報工学分野）

説明変数	2	3	4	5	6
1	41.51	51.89	49.06	42.45	50.00
2		50.00	47.17	43.40	42.45
3			52.83	51.89	50.94
4				47.17	43.40
5					40.57

表 3-5-3　判別率（コーパス：全データ）

説明変数	2	3	4	5	6
1	53.77	52.83	48.11	51.89	51.89
2		54.72	55.66	47.17	50.94
3			53.77	45.28	50.94
4				36.79	40.57
5					30.19

表 3-6　判別率

（説明変数は文字長、誤り個数、誤り文字長、文あたりの確率）

種類	判別率
1)標準	55.7
2)情報工学分野	51.9
3)医療分野	49.1
1)+2)	52.8
1)+3)	49.1
1)+2)+3)	54.7

第４章　まとめ

本研究では、Ｎ－ＧＲＡＭを利用した翻訳文の選別精度を各種データを用いて検討を行った。その結果をまとめると次のようになる。

【意味距離同一値候補の選別（相対的予測）】

- ●相対的予測の判別精度は約７０％であった。これはランダムに選ぶ場合（約４０％）と比較すると３０ポイントも高く、意味距離同一値候補の選択には N-GRAM が有用な手法であることがわかった。
- ●英語の場合、文字種が限られているので文字 N-GRAM での判別は低い考えていたが、N-GRAM 次元を上げ、コーパス量を増やすことでそこそこの判別精度が得られた。(今回の実験では最高判別率は、8-NGRAM、18452 文で６６％であった。ランダム選択の３０％と比較すると３６ポイント高い)。
- ●文字と混合 N-GRAM を比較すると判別精度は文字の方が高い値を示す傾向が見られた。

【訓練前後の翻訳データ】

- ●一文に含まれる各文字毎の N-GRAM 確率値において、一定の閾値以下の部分をカウントする「誤り個数」と各文字の値を平均した「文の N-GRAM 確率値」などの判別精度に対する寄与度合いを比較すると、後者の寄与が大きいことがわかった。「誤り個数」は悪い部分だけを注目するのに対し、「文の N-GRAM 確率値」は N-GRAM 確率の高い、滑らかな文字列をも評価しているためであろう。

【科学文献抄録翻訳データ】

- ●検証データの偏りなどで特にコメントできる結果を得られなかった。

翻訳文の評価を人間が行う場合、原言語と目的言語を比較して原言語にある単語が正しく訳されているか、原言語にある関係（修飾、並列関係など）を正しく捉えているか、等を確認している。一方、今回用いた N-GRAM は、目的言語だけを対象に、言語としての良否を判別してるだけなので、誤った訳であっても目的言語として正しければ良い得点を付けてしまう可能性がある。

付録　N-GRAM 作成、及び計算方法

1．概要

実験で用いた N-GRAM の作成方法および計算方法についてまとめた。N-GRAM の作成手順は次のとおりである。

（1）テキストの収集
（2）テキストの加工
（3）コード付け
（4）N連鎖単位の抽出と頻度のカウント
（5）N-GRAM 確率値ﾃｰﾌﾞﾙの作成
（6）入力テキストの N-GRAM 値の計算

2．テキストの収集

ATR のＳＬＤＢ、ＬＤＢの ETEXT、JTEXT からテキストを収集した（図 A-1 参照）。収集したテキストは後続処理の前準備として以下の加工を行った。加工後の結果を図 A-2 に示す。**本文におけるコーパスの文章数の単位は、この段階で作成した一テキストである。一テキストには複数の文が含まれる場合もある。**

●テキストに含まれる、[] や<>などの補助説明を除去する。
●英語の場合、単語の区切りである空白を一文字分にする。
●担当者、通訳者などを除去する。

customer:　[ah] Yes, (that's) that's right.
interpreter: Well, it's about ten minutes on foot from Keihan Sanjo station to get there. But [uh] we have a map at this front desk. So (if you kin) if you kindly come down (to) here, I can (ef) explain better.

担当者：[ああ] そうですか。それでは今夜九時までに出していただきますと、翌朝<yokuchou>八時までには上がると思いますよ。
通訳者：分かりました。[あ] それで結構です。
　　服はフロントに持って行けばいいんですか。
担当者：そうですね。フロントのわきにカウンターが有りますので、そこまでお持ちください。
　　それと料金が普通の料金の三十パーセント増しになりますが、それでよろしいですか。

図 A-1　ETEXT,JTEXT の例

Yes, that's right.
Well, it's about ten minutes on foot from Keihan Sanjo station to get there. But we have a map at this front desk. So if you kindly come down here, I can explain better.

そうですか。それでは今夜九時までに出していただきますと、翌朝八時までには上がると思いますよ。
分かりました。それで結構です。
服はフロントに持って行けばいいんですか。
そうですね。フロントのわきにカウンターが有りますので、そこまでお持ちください。
それと料金が普通の料金の三十パーセント増しになりますが、それでよろしいですか。

図 A-2　ETEXT,JTEXT の加工例

3．テキストの加工

N-GRAM は文字と品詞・単語混合の二種類を作成する。

日本語テキストは、まず最初に JUMAN による形態素解析を行う（図 A-3 参照)。文字 N-GRAM の場合は、この形態素結果をもとに数字や固有名詞を記号で置き換える文字の抽象化を行った（表 A-1、図 A-4 参照)。

表 A-1　日本語文字の抽象化

品詞	記号
数詞	@N
名詞の人名、地名、組織名、固有名詞	@K
未定義語	@S

英語テキストは、混合 N-GRAM 作成のために、Black　Tagger を用いてタグ付けを行う（図 A-5 参照)。

```
服はフロントに持って行けばいいんですか。
0 2 服 ふく 服 名詞 6 普通名詞 1 * 0 * 0 NIL
2 4 は は は 助詞 9 副助詞 2 * 0 * 0 NIL
4 12 フロント フロント フロント 名詞 6 普通名詞 1 * 0 * 0 NIL
12 14 に に に 助詞 9 格助詞 1 * 0 * 0 NIL
14 26 持って行けば もっていけば 持って行く 動詞 2 * 0 子音動詞カ行促音便形 3 基本条件形 6 NIL
26 30 いい いい いい 形容詞 3 * 0 イ形容詞アウオ段 18 基本形 2 NIL
30 36 んです んです んだ 助動詞 5 * 0 ナ形容詞 21 デス列基本形 26 NIL
36 38 か か か 助詞 9 終助詞 4 * 0 * 0 NIL
38 40 。 。 。 特殊 1 句点 1 * 0 * 0 NIL
EOS
```

図 A-3　JUMAN 解析例

```
そうですか、@N 日はちょっと都合がつかないので、じゃあ@N 日のビー席を@N 枚、お願いし
ます。
日にちが@N 月@N 日の土曜日、時間は@N 時半ぐらいからお願いします。
はい、@K@K 様となっております。
```

図 A-4　文字抽象化の例

```
0.0753428 [start Yes_NNS ._. we_PRP are_VBP right_JJ in_IN front_NN of_IN the_DT
 Hachijo_NNP Exit_NNP ._.  start]
0.264653 [start Yes._NNP That_DT will_MD be_VB possible._NN But_CC you_PRP 'll_M
D have_VB to_TO pay_VB additional_JJ charge_NN ._.  start]
```

図 A-5　Black　Tagger 解析例

4．コード付け

加工したデータを用いて、N-GRAM 単位のコード付けを行う。この時、テキストの文頭及び文末に開始、終了を表すダミーコードを挿入して、コーパスすべてのテキスト（コード化している）を一列に並べたデータを作成する（図 A-6 参照)。

```
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
06010000
08000000
04000024
88000001
01010000
10000000
13010000
06010000
04000025
88000002
01010000
90000002          .... 終了コード
90000001          .... 開始コード      次のテキスト
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
90000001          .... 開始コード
06010000
08000000
04000024
88000001
01010000
10000000
13010000
06010000
04000025
88000002
01010000
90000002          .... 終了コード
```

図 A-6　コード付け例

5．N連鎖単位の抽出と頻度のカウント

図 A-6 に示したデータをもとに、N連鎖する単位とその出現頻度をカウントする。

6．N-GRAM 確率値ﾃｰﾌﾞﾙの作成

N連鎖とN－1連鎖する単位の出現頻度から下式によって N-GRAM 確率値ﾃｰﾌﾞﾙを作成した。

$$\text{N-GRAM 確率値} = \mathrm{LOG}_{10}(\text{N連鎖頻度} / \text{N－1連鎖頻度})$$

7．入力テキストのN-GRAM値の計算

N-GRAMの計算は次の手順で行った。まず、入力テキストに対し、適用するN-GRAMと同じ加工を行いコード付けする（文頭、文末にはダミーコードを入れる）。そして、文頭の最初の単位から初めて、N-GRAM確率値ﾃｰﾌﾞﾙを参照しながら各単位ごとのN-GRAM確率値を求めた。