機械翻訳において MRR を用いた複数出力による自動評価方法

小川知紘 [*1] 村上仁一 [*2] 徳久雅人 [*2] 江木孝史 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*1,*2]{s082010, murakami, tokuhisas, s082008} @ ike.tottori-u.ac.jp

1　はじめに

現在，機械翻訳の翻訳品質の評価において, 複数の自動評価方法が提案されている．多くの評価方法は，翻訳された最尤の出力文と参照文から単語の順列や出現頻度を見て評価を行う．つまり，出力文 1 文に対して評価を行っている [1]．しかし，実際に翻訳を行う際には，複数の出力文の中から最適な文を選び，翻訳を行うことがある．

一方，情報検索においては最尤の文だけでなく，複数の文を使用し検索精度を評価する．評価指標の一つとして MRR がある．MRR は，検索課題毎に正解が出現した順位の逆数を求め (RR)，全検索課題の RR を平均することで，システムの検索精度を評価する．

そこで，本研究では，MRR を用いた複数の出力文による自動評価方法について調査する．

2　MRR

MRR とは 1 つの正解について正解が出現した順位を評価する指標である．検索課題毎に正解が最初に出現した順位の逆数を求め (RR(Reciprocal Rank)), 全検索課題の RR を平均すること (MRR) で, システムの検索精度を評価する．
計算式を以下に示す．

r：正解が出現した順位

N：全検索課題数

$$RRi = \frac{1}{r} \tag{1}$$

$$MRR = \frac{1}{N}\sum_{i}^{N} RRi \tag{2}$$

3　機械翻訳における自動評価

統計翻訳が広まるにつれて，自動評価も様々な方法が提案されてきた．本研究では，BLEU[1]，METEOR[2]，RIBES[3]，TER[4]，STR[5] を使用する．

STR(Sentence Translation Ratio) は，出力文と参照文の文一致数で評価を行う評価法である．

4　機械翻訳における人手評価 [6]

人手評価には，多くの手法があるが，その中でも Fluency と，Adequacy がよく利用されている．Fluency は言語としての流暢さから評価を行う．Adequacy は内容としての適切さから評価を行う．

本研究では，Adequacy を利用する．入力文と出力文を比較し，翻訳品質を 1 から 5 の数値でランクづけする手法で行う．ランク 1 がもっとも悪い評価で，ランク 5 がもっとも良い評価である．評価基準の例を表 1 に，評価の例を参照文とともに表 2 から表 5 に示す．本研究での評価者は一人である．

表 1　人手評価の評価基準

ランク	ランク付与基準
5	文法が正しく，言いたいことがすぐわかる． ネイティブレベルの文で，重要な情報の欠落はない．
4	文法が正しく，言いたいことがすぐわかる． 重要な情報の欠落はない．
3	文法に誤りがあるが，言いたいことがすぐわかる． 重要な情報の欠落はない．
2	文法に誤りがあるが，言いたいことがすぐわからない． 重要な情報の欠落はない．
1	文法に誤りがあり，言いたいことも分からない． 重要な情報が欠落している．

表 2　単文 日英翻訳 人手評価の例

	評価文	人手評価
入力文	この ボール は よく はずむ 。	5
出力文	The ball bounces well .	
参照文	This ball bounces well .	
入力文	下水 が よく 流れ ない 。	4
出力文	The drain is not flow .	
参照文	The drain does not flow well .	
入力文	彼女 は 古い 着物 を はやり の 色 に 染め直した 。	3
出力文	She redyed the old kimono a popular color .	
参照文	She redyed an old kimono in a color that is in fashion .	
入力文	日本 は 毎年 6 月頃 よく 雨 が 降る 。	2
出力文	In Japan are I often every year around June rains .	
参照文	It rains a lot in Japan around June of each year .	
入力文	その 機械 の 使い方 が だれ に も わから なかっ た 。	1
出力文	Nobody knew how to use the machine .	
参照文	No one was able to guess how to use the machine .	

表 3　単文 英日翻訳 人手評価の例

	評価文	人手評価
入力文	A carpenter made a wooden desk .	5
出力文	大工 は 木製 の 机 を 作っ た 。	
参照文	大工 が 木 の 机 を 作っ た 。	
入力文	The lecture was very boring and I slept through most of it .	4
出力文	講義 は とても 退屈 なので 私 は その 大部分 で 眠った 。	
参照文	その 講義 は たいへん 退屈 な もの だっ た ので 私 は その 大半 を 眠っ て い た 。	
入力文	It has gotten dark outside .	3
出力文	外 は 暗い なった 。	
参照文	外 が 暗く な っ た 。	
入力文	My bicycle chain came off on my way to school .	2
出力文	私 の 自転車 が 鎖 登校 の 途中 で 飛び立っ た 。	
参照文	登校 の 途中 で 自転車 の チェーン が 外れた 。	
入力文	Civil aviation in Japan is very backward .	1
出力文	後方 に は 日本 の だ 。	
参照文	日本 の 民間 飛行 は さっぱり 振 わ ない 。	

表 4 重文複文 日英翻訳 人手評価の例

	評価文	人手評価
入力文	大阪 へ 向かう 列車 の 中 で 友達 に 会っ た 。	5
出力文	I met a friend in the train to Osaka .	
参照文	I ran into a friend on the train for Osaka .	
入力文	彼 は 信念 の ある 記事 を 書く 。	4
出力文	He writes an account of belief .	
参照文	He writes edgy articles .	
入力文	その 材料 を 組み合わせ て きわめて 軽い 飛行機 を 作る 。	3
出力文	The materials are together and made an extremely light airplane .	
参照文	Those materials are combined to create airplanes that are extremely light.	
入力文	人間 と 獣 を 分け て いる の は 知性 で ある 。	2
出力文	It is intelligence to the beasts man .	
参照文	It is intelligence that differentiates man from the beasts .	
入力文	針 一 本 落ち て も 聞こえ そう だ 。	1
出力文	A the sound of a book .	
参照文	A pin might be heard to drop .	

表 5 重文複文 英日翻訳 人手評価の例

	評価文	人手評価
入力文	I cannot believe him any longer	5
出力文	もう これ 以上 彼 の 言う 事 は 信じ られ ない 。	
参照文	あの 人 の 言う こと は もう 本当に は でき ない 。	
入力文	A clown wobbled past on a unicycle .	4
出力文	ピエロ は 、 一輪車 で 通り過ぎ て いっ た 。	
参照文	道化師 が 一輪車 に 乗っ て よろよろ 通り過ぎ た 。	
入力文	He is wanting in that knowledge which is requisite to a teacher .	3
出力文	彼 は 先生 に は 必要 な 知識 が に は 足りない 。	
参照文	彼 は 教師 たる もの に 必要 な 知識 を 欠い て いる 。	
入力文	It is bad for health to exercise hard right after a meal .	2
出力文	それ は 身体 に 悪い 食事 を した 後 に 右 の 運動 を し て いた 。	
参照文	食後 すぐ に 激しい 運動 を する の は 体 に 良く ない 。	
入力文	Out of five partners , one dropped out , leaving four .	1
出力文	5 人 が 共同 して 、 4 落とし た 。	
参照文	5 人 の 仲間 から ひとり 抜け て 4 人 に なっ た 。	

5 提案手法

本研究では，複数出力された文を使用して評価を行う．具体的には入力文 1 文につき，最尤の出力文から上位 4 文の出力文を得る．次に，それぞれの文に自動評価を行い *MRR* を用いて評価値を求める．概略を図 1 に示す．

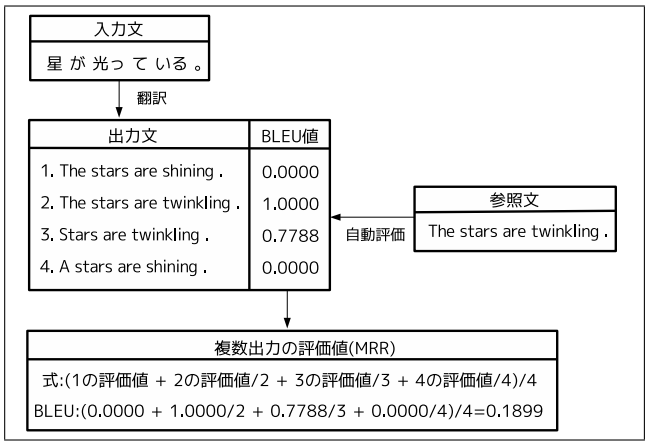

図 1 提案手法の手順

6 実験

6.1 翻訳システム

本研究では，7 種類の翻訳システムを使用する．使用する翻訳システムを表 6 に示す．以下に翻訳システムについて述べる．

6.1.1 ルールベース翻訳

ルールベース翻訳とは，人手によって構成された変換規則を元に翻訳を行うシステムである．本研究では，東芝の Tauras[7] を使用する．

6.1.2 句に基づく統計翻訳

本研究では，moses[8] を使用する．

6.1.3 階層型統計翻訳

本研究では，moses[8] を使用する．

6.1.4 ハイブリッド翻訳

ハイブリッド翻訳は，前処理にルールベース翻訳を用いる．そして後処理に統計翻訳を使用する翻訳システムである [9]．ルールベース翻訳には，東芝の Tauras[7] と富士通の Atras[10] を使用する．

表 6 翻訳システム

翻訳システム	略記
Tauras を使用したルールベース翻訳	RBMT
句に基づく統計翻訳	PSMT
階層型統計翻訳	HSMT
前処理に Tauras を使用したルールベース翻訳を用いた，句に基づく統計翻訳 (ハイブリッド翻訳)	RBMT(t)+PSMT
前処理に Atras を使用したルールベース翻訳を用いた，句に基づく統計翻訳 (ハイブリッド翻訳)	RBMT(a)+PSMT
前処理に Tauras を使用したルールベース翻訳を用いた，階層型統計翻訳 (ハイブリッド翻訳)	RBMT(t)+HSMT
前処理に Atras を使用したルールベース翻訳を用いた，階層型統計翻訳 (ハイブリッド翻訳)	RBMT(a)+HSMT

6.2 実験データ

本実験では，以下の 2 種類のコーパスを使用する．

6.2.1 単文コーパス [11]

単文コーパスは，日本語が単文である対訳コーパスである．コーパスの文は辞書の例文から抽出している．本実験では学習データとして 100,000 文，ディベロップメントデータとして 1,000 文を用いる．日英翻訳にはテスト文として 3,744 文，英日翻訳にはテスト文として 1,122 文を用いる．単文コーパスの例を表 7 に示す．

表 7 単文コーパスの例

日本語文	コンピューター は 2 進 法 の 2つ の 数 を 用いる 。
英語文	A computer employs the two digits of the binary system .

6.2.2 重文複文コーパス [11]

重文複文コーパスは，日本語が重文複文である対訳コーパスである．コーパスの文は，辞書の例文から抽出している．本実験では，学習データとして 100,000 文，ディベロップメントデータとして 1,000 文を用いる．日英翻訳にはテスト文として 5,435 文，英日翻訳にはテスト文として 1,280 文を用いる．重文複文コーパスの例を表 8 に示す．

表 8 重文複文コーパスの例

日本語文	腹 の 皮 が よじれる ほど 笑っ た 。
英語文	I almost split my sides laughing .

6.3 翻訳の種類

本実験では，日英翻訳と英日翻訳の 2 種類について行う．

6.4 評価方法

評価は，複数文出力の実験と 1 文出力の実験の 2 種類で行う．1 つのテスト文に対し, 最尤の出力文から上位 4 文ずつ出力文を得る．次にそれぞれの文に対し自動評価を行い *MRR* を用いて評価値を求める．

7 実験結果

実験は，翻訳システム 7 種類，実験データ 2 種類，翻訳の種類 2 種類，評価方法 2 種類の計 56 種類の実験を行う．実験結果を表 9，表 10，表 12，表 13，表 15，表 16，表 18，表 19 に示す．人手評価との相関係数を表 11，表 14，表 17，表 20 に示す．

表 9 単文 日英翻訳 1 文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.1196	0.4500	0.7123	0.7611	0.0075	**3.34**
PSMT	0.1242	0.4377	0.6910	0.7433	0.0091	2.01
HSMT	0.1247	0.4429	0.6971	0.7478	0.0075	2.13
RBMT(a) +PSMT	**0.1634**	**0.4915**	0.7309	**0.7003**	**0.0272**	2.81
RBMT(a) +HSMT	0.1605	0.4877	**0.7338**	0.7016	0.0254	2.79
RBMT(t) +PSMT	0.1507	0.4772	0.7300	0.7127	0.0155	3.03
RBMT(t) +HSMT	0.1489	0.4776	0.7290	0.7147	0.0147	3.04

表 10 単文 日英翻訳 複数文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.0480	0.2074	0.3532	0.4458	0.076	**1.6638**
PSMT	0.0626	0.2263	0.3592	0.3902	0.0120	1.0525
HSMT	0.0630	0.2292	0.3619	0.3921	0.0114	1.1127
RBMT(a) +PSMT	0.0814	**0.2530**	0.3795	0.3675	**0.0326**	1.4602
RBMT(a) +HSMT	**0.0825**	0.2529	**0.3813**	**0.3669**	0.0304	1.4606
RBMT(t) +PSMT	0.0757	0.2462	0.3792	0.3736	0.0193	1.5635
RBMT(t) +HSMT	0.0768	0.2482	0.3790	0.3728	0.0172	1.5710

表 11 単文 日英翻訳 人手評価との相関係数

	BLEU	METEOR	RIBES	TER	STR
人手評価 (1 文出力)	0.3150	0.5305	0.7203	-0.2108	0.2402
人手評価 (複数文出力)	0.1300	0.1410	0.3492	0.1497	0.2120

表 12 単文 英日翻訳 1 文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.1185	0.3693	0.7371	0.9059	0.0027	**3.08**
PSMT	0.1517	0.4093	0.7397	0.8240	0.0134	2.21
HSMT	0.1585	0.4163	0.7511	0.8161	0.0169	2.23
RBMT(a) +PSMT	0.1975	0.4702	0.7653	0.7581	**0.0258**	2.97
RBMT(a) +HSMT	**0.2003**	**0.4717**	**0.7784**	**0.7503**	0.0250	3.00
RBMT(t) +PSMT	0.1878	0.4599	0.7679	0.7613	0.0223	3.00
RBMT(t) +HSMT	0.3999	0.4649	0.7729	0.7566	0.0223	2.92

表 13 単文 英日翻訳 複数文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.0525	0.1768	0.3738	0.4972	0.0036	1.4896
PSMT	0.0755	0.2106	0.3835	0.4328	0.0173	1.1392
HSMT	0.0804	0.2157	0.3896	0.4263	0.0217	1.1729
RBMT(a) +PSMT	0.1012	0.2436	0.3977	0.3967	**0.0305**	1.5046
RBMT(a) +HSMT	**0.1042**	**0.2451**	**0.4055**	**0.3909**	0.0274	**1.5627**
RBMT(t) +PSMT	0.0949	0.2370	0.3981	0.3985	0.0262	1.5315
RBMT(t) +HSMT	0.1000	0.2409	0.4026	0.3952	0.0251	1.5206

表 14 単文 英日翻訳 人手評価との相関係数

	BLEU	METEOR	RIBES	TER	STR
人手評価 (1 文出力)	0.2953	0.3293	0.4846	0.1853	0.1713
人手評価 (複数文出力)	0.4107	0.3833	0.4820	0.2637	0.2437

表 15 重文複文 日英翻訳 1 文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.0914	0.3930	0.6653	0.8565	0.0015	**3.00**
PSMT	0.1138	0.4058	0.6650	0.7877	0.0042	2.25
HSMT	0.1166	0.4192	0.67052	0.7755	0.0035	2.32
RBMT(a) +PSMT	0.1325	0.4370	0.6769	0.7588	**0.0090**	2.52
RBMT(a) +HSMT	0.1346	0.4371	0.6839	0.7591	**0.0090**	2.47
RBMT(t) +PSMT	**0.1395**	**0.4455**	0.7010	0.7501	0.0059	2.54
RBMT(t) +HSMT	0.1366	0.4405	**0.7026**	**0.7480**	0.0057	2.54

表 16 重文複文 日英翻訳 複数文出力

	BLEU	METEOR	RIBES	TER	MRR	人手評価
RBMT	0.0390	0.1834	0.3334	0.4851	0.0015	**1.5190**
PSMT	0.0576	0.2101	0.3453	0.4130	0.0052	1.1690
HSMT	0.0595	0.2171	0.3488	0.4056	0.0044	1.2075
RBMT(a) +PSMT	0.0674	0.2261	0.3511	0.3975	**0.0108**	1.3098
RBMT(a) +HSMT	0.0687	0.2268	0.3549	0.3969	0.0103	1.3048
RBMT(t) +PSMT	**0.0713**	**0.2310**	0.3642	0.3968	0.0080	1.3189
RBMT(t) +HSMT	0.0705	0.2288	**0.3656**	**0.3907**	0.0074	1.3235

表 17 重文複文 日英翻訳 人手評価との相関係数

	BLEU	METEOR	RIBES	TER	STR
人手評価 (1 文出力)	-0.4552	-0.3099	-0.0196	0.6455	-0.3364
人手評価 (複数文出力)	-0.4942	-0.5202	-0.3168	0.7027	-0.2863

表 18 重文複文 英日翻訳 1 文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.1130	0.3717	0.6998	0.8979	0.0008	**3.05**
PSMT	0.1407	0.3875	0.6985	0.8597	0.0055	2.16
HSMT	0.1417	0.3974	0.7070	0.8406	0.0047	2.24
RBMT(a) +PSMT	**0.1900**	**0.4559**	0.7472	0.7709	**0.0094**	2.80
RBMT(a) +HSMT	0.1884	0.4518	**0.7571**	**0.7698**	0.0047	2.73
RBMT(t) +PSMT	0.1823	0.4499	0.7491	0.7780	0.0063	2.75
RBMT(t) +HSMT	0.1791	0.4488	0.7500	0.7789	0.0047	2.82

表 19 重文複文 英日翻訳 複数文出力

	BLEU	METEOR	RIBES	TER	STR	人手評価
RBMT	0.0517	0.1810	0.3562	0.4877	0.0008	**1.5252**
PSMT	0.0712	0.1997	0.3624	0.4500	0.0074	1.1330
HSMT	0.0732	0.2056	0.3676	0.4395	0.0049	1.1723
RBMT(a) +PSMT	**0.0971**	**0.2359**	0.3889	**0.4033**	**0.0116**	1.4575
RBMT(a) +HSMT	0.0958	0.2329	**0.3933**	0.4037	0.0070	1.4202
RBMT(t) +PSMT	0.0937	0.2336	0.3895	0.4062	0.0081	1.4180
RBMT(t) +HSMT	0.0925	0.2329	0.3904	0.4073	0.0057	1.4629

表 20 重文複文 英日翻訳 人手評価との相関係数

	BLEU	METEOR	RIBES	TER	STR
人手評価 (1 文出力)	0.1768	0.3070	0.4405	-0.1906	-0.2184
人手評価 (複数文出力)	0.1981	0.2744	0.4048	-0.1546	-0.1016

RBMT と RBMT(a)+HSMT の英日翻訳の単文と重文複文の複数出力文の例を以下に示す.

表 21 単文 英日翻訳 複数出力文の例

入力文	The wound requires prompt treatment .
参照文	この 傷 は 至急 手当 を せ ね ば なら ない 。
出力文 (RBMT)	傷 は 迅速 な 処理 を 要求 する 。
	傷 は プロンプト 処理 を 要求 する 。
	巻か れ た もの は 迅速 な 処理 を 要求 する 。
	巻か れ た もの は プロンプト 処理 を 要求 する 。
出力文 (RBMT(a) +HSMT)	傷 は 治療 を 必要 と し ます 。
	傷口 が 治療 を 必要 と し ます 。
	傷口 が 素早い 治療 を 必要 と し て いる 。
	傷口 は 治療 を 必要 と し ます 。

表 22 重文複文 英日翻訳 複数出力文の例

入力文	He is wanting in that knowledge which is requisite to a teacher .
参照文	彼 は 教師 たる もの に 必要 な 知識 を 欠い て いる 。
出力文 (RBMT)	彼は、教師に必要なその知識に欠けている。
	彼は、教師への必要条件であるその知識に欠けている。
	彼は、教師に必要なその知識の中で望んでいる。
	彼は、教師への必要条件であるその知識の中で望んでいる。
出力文 (RBMT(a) +HSMT)	彼 は あの 先生 に 必要 な 知識 を 欠い て いる 。
	彼 は その 先生 に 必要 な 知識 を 欠い て いる 。
	彼 は 先生 に 必要 な 知識 を 欠い て いる 。
	彼 は 教師 として 必要 な 知識 を 欠い て いる 。

表 12 より, 単文の英日翻訳において, 複数文出力と 1 文出力を比較すると人手評価との相関係数が上がった. しかし, 他の実験では複数文出力と 1 文出力を比較しても, 人手評価との相関係数に差はあまり見られなかった.

8 考察

単文の英日翻訳においては, 複数文出力と 1 文出力を比較すると人手評価との相関係数が上がった. しかし, 他の実験では複数文出力と 1 文出力を比較しても人手評価との相関係数に差はあまり見られなかった.

この原因として, 単文の英日翻訳の 1 文出力では, RBMT の人手評価の値が最も高いが, 複数文出力では, RBMT(a)+HSMT の人手評価の値が最も高くなっている. そのため, 複数文出力の自動評価が 1 文出力の自動評価に比べ人手評価との相関係数が高くなったと考えている.

しかし, 重文複文の英日翻訳の実験では, 複数文出力と 1 文出力を比べても人手評価の相関係数にあまり差は見られない. また, 人手評価は複数文出力も 1 文出力も RBMT が最も高くなっている. 単文と重文複文で違いが出たのは人手評価に原因があるとも考えられる. 今後, 評価者を増加させて, 人手評価の信頼性を高める必要がある.

9 おわりに

本研究では, *MRR* を用いた複数の出力文による自動評価方法について調査した. しかし, 複数文出力の実験と 1 文出力の実験において, 人手評価との相関係数にはあまり変化が見られなかった. 今後は, 人手評価の再調査を行っていきたい. また, ルールベース翻訳が複数文を出力する数が少なかったため, テスト文の数も少なくなった. テスト文の数を増やした調査も必要である.

参考文献

[1] Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu: “BLEU: a method for automatic evaluation of machine translation”, 40th Annual meeting of the Association for Computational Linguistics, pp.311–318, 2002.

[2] Banerjee Satanjeev, Lavie Alon: “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”, Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), pp.65–72, 2005.

[3] 平尾努, 磯崎秀樹, Kevin Duh, 須藤克仁, 塚田元, 永田昌明: “RIBES: 順位相関に基づく翻訳の自動評価法”, 言語処理学会第 17 年次大会発表論文集, pp.1111–1114, 2011.

[4] Richard Schwartz, Linnea Micciulla, John Makhoul. “A Study of Translation Edit Rate with Targeted Human Annotation”, AMTA, 2006.

[5] 石原 雅文, ” 文一致数を用いた機械翻訳の自動評価”, 平成 24 年度卒業論文,2013.

[6] 奈良先端科学技術大学院大学 情報科学研究科, Graham Neubig, “文レベルの機械翻訳評価尺度に関する調査”

[7] Shinya Amano, Hideki Hirakawa, Yoshinao Tsutsumi, “TAURAS: The Toshiba machine translation system”, Manuscr Program MT Summit, pp.15–23, 1987.

[8] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, June 2007.

[9] 村上仁一, 徳久雅人, “ルールベース翻訳と統計翻訳を結合した特許翻訳”, AAMT/Japio 特許翻訳研究会 第 1 回特許情報シンポジウム, pp.46-53, Dec. 2010.

[10] 英 日・日 英 翻 訳 ソ フ ト ATLAS: http://software.fujitsu.com/jp/atlas/

[11] 第一回コーパス日本語学ワークショップ, 村上仁一, 藤波進, “日本語と英語の対訳文対の収集と著作権の考察.pdf ”, pp.119-130, Mar. 2012.