未知語処理のための頻度1単語の精度調査

初田直樹 [*1]　村上仁一 [*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
s132044@ike.tottori-u.ac.jp[*1] murakami@eesc.tottori-u.ac.jp[*2]

1　はじめに

機械翻訳において，入力文中の単語が未知語として出現する．この未知語を翻訳する手法（以下，未知語処理）が，数多く提案されている [1]．しかし，未知語処理にはまだ課題が多い．

未知語処理の 1 つとして，対訳学習文と対訳単語確率をもとに作成した対訳単語辞書を用いる手法がある [1]．そして，対訳単語辞書の精度を調査する研究がある [2]．中村らの研究 [2] では，対訳単語辞書の日本語単語と英語単語の適切な対応の数を調査した．この研究の問題点は，主に 2 点ある．1 点目は，対訳単語確率の計算において IBM 翻訳モデル [3] の Model1 のみの調査で，他のモデルでの調査が行われていない．2 点目は，全ての対訳単語を調査対象としており，未知語処理を対象としていない．対訳単語確率を計算するモデルとしては，IBM Model1〜Model5 や HMM[4] などがある．また未知語の多くは，対訳学習文中に 1 回のみ出現する単語（以下，頻度 1 単語）である．

そこで本研究では，IBM Model1〜Model5 と HMM の各モデルにおいて，頻度 1 単語の翻訳精度を調査する．

2　未知語処理

機械翻訳には，未知語処理の手法が数多く提案されている．本章では，本論文で扱う未知語処理の流れ [1] を図 1 に示す．また，具体的な手順を以下に示す．

手順 1　未知語を含む文の生成
任意の機械翻訳機より，未知語を含む文を出力する．

表 1　未知語を含む出力文

入力文	彼は誤植を見付けた.
出力文	He found a 誤植.

手順 2　未知語を翻訳
手順 1 で出力された未知語を対訳単語辞書を用いて翻訳する．訳語が複数存在する場合は，対訳単語確率が最大のものを選択する．なお，表 2 の $\log_2(P(E|J))$ は，日本語単語が英語単語に訳される GIZA++[5] の対訳単語確率である．

表 2　対訳単語辞書

日本語	英語	$\log_2(P(E\|J))$
誤植	misprint	-3.816

手順 3　未知語処理後の出力文の生成
未知語処理後の出力文を生成する．

表 3　未知語処理後の出力文

入力文	彼は誤植を見付けた.
出力文	He found a misprint.

3　対訳単語辞書

2 章の手順 2 で使用している対訳単語辞書は，対訳学習文と GIZA++ を用いて，対訳単語に対訳単語確率を付与して作成する [2]．対訳単語辞書の作成の流れを図 2 に示す．また，具体的な手順を以下に示す．

手順 1　単語対応の取得
対訳学習文と GIZA++ から，日英方向の単語対応と英日方向の単語対応を取得する．なお，表 4 の $\log_2(P(J|E))$ は，英語単語が日本語単語に訳される GIZA++ の対訳単語確率である．

表 4　単語対応の取得

日本語単語	英語単語	$\log_2(P(E\|J))$	$\log_2(P(J\|E))$
本	book	-0.297	-0.144
誤植	misprint	-3.816	-1.824

手順 2　枝刈り処理
手順 1 で作成した対訳単語辞書には，明らかに不適切な対訳単語が多数含まれている．そこで，任意の条件（以下，枝刈り条件）を満たす対訳単語のみに限定する処理を行う．

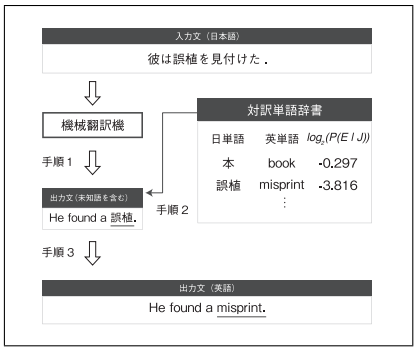

図 1　未知語処理の流れ図

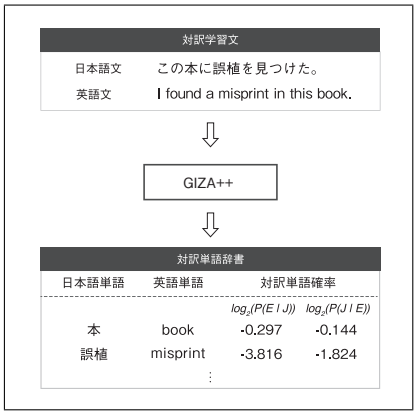

図 2　対訳単語辞書作成の流れ図

4 従来研究の問題点

従来研究 [2] は，対訳単語辞書の日本語単語と英語単語の適切な対応の数を調査した．この研究の問題点を以下に示す．

1. 対訳単語確率の計算において，IBM Model1 のみの調査であり，IBM Model2〜Model5 での調査は行われていない．
2. 全ての対訳単語を調査対象としており，未知語処理を対象とした調査ではない．

5 調査手法

5.1 調査対象

本論文では，GIZA++ の対訳単語確率の計算において，IBM Model1〜Model5 と HMM を用いる．また，未知語の多くが頻度 1 単語であることから，頻度 1 単語の翻訳精度を調査する．なお，本論文の頻度 1 単語は，対訳学習文中の日本語単語と英語単語が共に出現する頻度（共起頻度）が 1 回の単語を指す．

5.2 対訳単語確率の計算方法

IBM 翻訳モデル [3] と HMM[4] の概要を以下に示す．

Model1 目的言語の単語が原言語の単語に訳される確率
Model2 Model1 + 単語の位置（絶対位置）を考慮
Model3 Model2 + 1 つの単語が複数対応
Model4 Model3 + 各単語の位置（相対位置）を考慮
Model5 Model4 + 単語の位置が重複する問題を解消
HMM Model1 + 単語の位置（相対位置）を考慮

6 実験

6.1 調査条件

調査は，電子辞書などの例文より抽出した単文コーパス [6] を用いる．使用するデータの内訳を表 5 に示す．

表 5 使用するデータ

対訳学習文	160,000 文

調査は, IBM Model1〜Model5 の 5 つのモデル（Model1〜Model3 は，HMM と組み合わせる）で行う．各モデルの GIZA++ のパラメータを以下に示す．なお，パラメータの左辺はモデルを表し，右辺の数値は学習回数を表す．

Model1 m1=4, m2=0, mh=4, m3=0, m4=0, m5=0
Model2 m1=0, m2=4, mh=4, m3=0, m4=0, m5=0
Model3 m1=0, m2=0, mh=4, m3=4, m4=0, m5=0
Model4 m1=0, m2=0, mh=0, m3=0, m4=4, m5=0
Model5 m1=0, m2=0, mh=0, m3=0, m4=0, m5=4

6.2 枝刈り条件

作成した対訳単語辞書において，以下の条件のみに限定する処理を行う．

- GIZA++ の対訳単語確率が，日本語・英語ともに $log_2(0.01)$ より高い単語
- GIZA++ の対訳単語確率を用いて作成した対訳単語の順位が，日本語・英語それぞれ 8 位以内の単語

6.3 評価方法

評価は全て，評価者 1 名で人手評価を行う．具体的には，対訳学習文と GIZA++ を用いて作成した対訳単語辞書から頻度 1 単語をランダムで 100 単語取り出して評価する．評価基準を表 6 に示す．

表 6 評価基準

○：適切な対訳単語
×：不適切な対訳単語

7 実験結果

7.1 評価結果

IBM Model1〜5 の頻度 1 単語の精度を調査した．評価結果を表 7 に示す．表中の全単語数は，作成した対訳単語辞書の全対訳単語数である．

表 7 評価結果

モデル	○	×	頻度 1 単語数	(全単語数)
Model1	**37**	**63**	52,271	86,857
Model2	32	68	51,757	86,401
Model3	33	67	46,083	76,679
Model4	32	68	48,575	82,838
Model5	34	64	47,980	82,136

頻度 1 単語は，全てのモデルで全単語数の約 60 %であった．つまり，頻度 1 単語が非常に多いことが分かる．翻訳精度は，Model1 がやや高かったが，どのモデルも大きな差はなかった．

7.2 評価例

次に，対訳単語辞書に含まれる日本語単語「備品」の全ての例を表 8 に示す．また，表中の各項目の概要を以下に示す．

- 日順： 対訳単語辞書の日本語単語を $\log_2(P(E|J))$ の順に並べた順位
- 英順： 対訳単語辞書の英語単語を $\log_2(P(J|E))$ の順に並べた順位
- 日頻： 対訳学習文中に出現する日本語単語の単語数
- 英頻： 対訳学習文中に出現する英語単語の単語数

表 8 対訳単語辞書に含まれる日本語単語「備品」の全ての例

モデル	日語 英語	評価	$\log_2(P(E\|J))$ $\log_2(P(J\|E))$	日順 英順	日頻 英頻
Model1	備品	×	-3.589	2	5
	requisition		-3.176	1	2
Model2	備品	×	-3.599	2	5
	requisition		-3.158	2	2
Model3	備品	○	-5.317	8	5
	fixtures		-2.450	2	1
Model3	備品	×	-3.735	2	5
	requisition		-3.347	2	2
Model4	備品	×	-4.250	3	5
	benches		-3.169	1	1
Model4	備品	○	-4.250	4	5
	fixtures		-2.493	2	1
Model4	備品	×	-4.250	2	5
	requisition		-3.315	2	2
Model5	備品	×	-4.181	2	5
	benches		-3.004	1	1
Model5	備品	○	-4.182	4	5
	fixtures		-2.487	2	1
Model5	備品	×	-4.181	3	5
	requisition		-3.301	2	2

表 8 の例では，Model3〜Model5 において，日本語単語「備品」に対して，英語単語「fixtures」という適切な対訳単語が含まれている．しかし，「requisition」や「benches」など，不適切な対訳単語も複数生成されている．

8 考察

8.1 作成した対訳単語辞書の全対訳単語の調査

表 7 では，頻度 1 単語の翻訳精度を調査した．本節では，全対訳単語で翻訳精度の調査を行う．評価方法は，全対訳単語からランダムで 100 語取り出して評価者 1 名で人手評価を行う．評価結果を表 9 に示す．

表 9　全対訳単語での評価結果

モデル	○	×	全単語数
Model1	55	45	86,857
Model2	53	47	86,401
Model3	**56**	**44**	76,679
Model4	54	46	82,838
Model5	52	48	82,136

表 9 より，全対訳単語においても翻訳精度に大きな差はなかった．また，表 7 と比較すると全てのモデルで翻訳精度が向上しており，頻度 1 単語の翻訳精度が低いことが分かる．

8.2 学習回数の変更

表 7 では，全てのモデルで学習回数を 4 回とした．本節では，IBM Model1 を用いて学習回数を 2,8,16 回と変更して調査を行う．GIZA++ のパラメータを以下に示す．

2 回　m1=2, m2=0, mh=2, m3=0, m4=0, m5=0
4 回　m1=4, m2=0, mh=4, m3=0, m4=0, m5=0
8 回　m1=8, m2=0, mh=8, m3=0, m4=0, m5=0
16 回　m1=16, m2=0, mh=16, m3=0, m4=0, m5=0

8.2.1 評価結果

各学習回数での評価結果を表 10 に示す．また，対訳単語辞書に含まれる日本語単語「備品」の全ての例を表 11 に示す．

表 10　学習回数を変更しての評価結果

学習回数	○	×	頻度 1 単語数	(全単語数)
2	31	69	48,050	79,797
4	**37**	**63**	52,271	86,857
8	31	69	52,213	87,720
16	33	67	51,896	87,452

表 11　対訳単語辞書に含まれる日本語単語「備品」の全ての例

学習回数	日語 英語	評価	$\log_2(P(E\|J))$ $\log_2(P(J\|E))$	日順 英順	日頻 英頻
2	備品	○	-4.929	8	5
	fixtures		-2.046	1	1
2	備品	×	-3.749	2	5
	requisition		-3.390	5	2
2	備品	×	-4.429	6	5
	swivel		-5.633	7	3
4	備品	×	-3.589	2	5
	requisition		-3.176	1	2
8	備品	×	-3.528	2	5
	requisition		-3.181	1	2
16	備品	×	-3.607	2	5
	requisition		-3.235	1	2

表 10 より，学習回数を変更しても翻訳精度に大きな差はなかった．

8.2.2 対訳単語の種類別の評価

各学習回数で翻訳精度の特徴を調べるために，対訳単語をひらがな，カタカナ，漢字と分けて評価した．対訳単語の種類別の評価を表 12 に示す．また，ひらがなとカタカナの対訳単語の例を表 13 に示す．

表 12　種類別の対訳単語の評価結果

学習回数	ひらがな	カタカナ	漢字
2	0.214 (3/14)	0.400 (6/15)	0.304 (21/69)
4	0.200 (3/15)	**0.615 (8/13)**	0.366 (26/71)
8	**0.250 (4/16)**	0.500 (5/10)	0.301 (22/73)
16	0.125 (2/16)	0.400 (4/10)	**0.369 (27/73)**

表 13　ひらがなとカタカナの対訳単語の例

学習回数	日語 英語	評価	$\log_2(P(E\|J))$ $\log_2(P(J\|E))$	日順 英順	日頻 英頻
2	ざわざわ	×	-4.504	6	5
	pines		-6.349	7	5
2	ざわざわ	○	-4.293	2	5
	soughed		-2.822	5	2
2	スロベニア	×	-2.320	2	1
	Elections		-4.163	3	2
2	スロベニア	○	-2.204	2	1
	Slovenia		-5.116	1	2

表 12 より，学習回数が 16 回の時は，他の学習回数の時よりひらがなの対訳単語の翻訳精度が低く，漢字の対訳単語の翻訳精度が高かった．この原因は，以下のように考えている．EM 推定においては，学習回数を増やしていくと翻訳候補が一意に絞られやすくなる傾向がある．そのため，一つの意味しか持たない漢字の対訳単語は，複数の意味を持つひらがなの対訳単語より，翻訳精度が高くなる．

8.3 順位の枝刈り条件の変更

8.3.1 日本語順位 1 位・英語順位 8 位以内

表 7 では，順位による枝刈り条件を日本語・英語ともに 8 位以内の対訳単語を調査対象とした．本項では，日本語順位 1 位のみ・英語順位 8 位以内の対訳単語で調査を行う．評価結果を表 14 に示す．また，対訳単語辞書に含まれる日本語単語「両生類」の全ての例を表 15 に示す．

表 14　日本語順位 1 位のみ・英語順位 8 位以内での評価結果

モデル	○	×	頻度 1 単語数	(全単語数)
Model1	45	55	13,834	28,277
Model2	41	59	13,860	28,308
Model3	**47**	**53**	10,373	23,170
Model4	35	65	16,426	31,470
Model5	45	55	16,082	30,999

表 15　対訳単語辞書に含まれる日本語単語「両生類」の全ての例

モデル	日語 英語	評価	$\log_2(P(E\|J))$ $\log_2(P(J\|E))$	日順 英順	日頻 英頻
Model1	両生類	×	-2.310	1	2
	semiterrestrial		-3.093	1	1
Model2	両生類	○	-2.366	1	2
	requisition		-2.795	1	1
Model3	両生類	○	-2.799	1	2
	requisition		-2.609	1	1
Model4	両生類	○	-3.200	1	2
	requisition		-2.804	1	1
Model5	両生類	×	-3.147	1	2
	semiterrestrial		-1.839	1	1

表 7 と表 14 を比較すると，全てのモデルで翻訳精度が向上した．この結果より，順位が高い対訳単語は翻訳精度が高いといえる．しかし，表 7 と表 14 の頻度 1 単語数を比べると，約 25 %に減少している．よって，日本語順位の枝刈り条件を 1 位のみに変更すると翻訳精度が向上するが，生成する対訳単語数が大幅に減少する．

8.3.2 日本語・英語ともに順位 1 位

8.3.1 項では，順位による枝刈り条件を日本語 1 位のみ・英語順位 8 位以内に変更して調査を行った．本項では，日本語・英語ともに順位が 1 位のみで調査を行う．評価結果を表 16 に示す．また，対訳単語辞書に含まれる日本語単語「叱責」の全ての例を表 17 に示す．

表 16　日本語・英語ともに順位が 1 位のみでの評価結果

モデル	○	×	頻度 1 単語数	(全単語数)
Model1	**52**	**48**	5,908	13,065
Model2	**52**	**48**	5,952	13,113
Model3	46	54	4,235	10,760
Model4	42	58	8,038	15,510
Model5	42	58	7,907	15,368

表 17　対訳単語辞書に含まれる日本語単語「叱責」の全ての例

モデル	日語 英語	評価	$\log_2(P(E\|J))$ $\log_2(P(J\|E))$	日順 英順	日頻 英頻
Model1	叱責	○	-3.921	1	4
	reprehension		-1.832	1	1
Model2	叱責	×	-3.951	1	4
	self-abuse		-2.431	1	1
Model3	叱責	○	-4.148	1	4
	reprehension		-2.214	1	1
Model4	叱責	×	-3.784	1	4
	self-abuse		-2.982	1	1
Model5	叱責	×	-4.054	1	4
	self-abuse		-2.791	1	1

表 14 と表 16 を比較すると，Model3 と Model5 では翻訳精度が低下した．よって，順位が日本語・英語ともに 1 位の対訳単語は，必ずしも適切ではないことが分かる．表 9 と比較しても，翻訳精度が低かった．つまり，頻度 1 単語は翻訳精度が非常に低いといえる．また，対訳単語数も表 14 と比べると約 50 %に減少している．

8.4 未知語処理の翻訳実験

8.4.1 実験条件

表 7 で使用した頻度 1 単語の対訳単語辞書を用いて，未知語処理の翻訳実験を行う．機械翻訳機には，ニューラル機械翻訳 [7] のツールキット OpenNMT[8] を使用した．使用データは，ニューラル機械翻訳 [7] を用いて得られた 1,000 文（そのうち未知語を含む文は 213 文）を入力文とした．

8.4.2 自動評価結果

自動評価は，入力文 1,000 文と各モデルの対訳単語辞書を用いて行った．未知語処理した（各モデルの対訳単語辞書を用いる）場合と，未知語処理しない（対訳単語辞書なし）場合の自動評価結果を表 18 に示す．

表 18　自動評価結果

モデル	BLEU	METEOR	RIBES	TER
Model1	0.200	**0.486**	0.783	**0.581**
Model2	0.200	**0.486**	0.783	**0.581**
Model3	0.200	**0.486**	0.783	**0.581**
Model4	0.200	**0.486**	**0.784**	**0.581**
Model5	**0.201**	**0.486**	0.783	**0.581**
なし	0.200	0.485	0.782	0.582

表 18 より，自動評価は全てのモデルでほとんど差がなかった．この結果の要因の 1 つは，未知語処理が正しく行われた場合でも，文全体の自動評価結果にはあまり影響しないためであると考えている．

8.4.3 人手評価結果

翻訳実験は，未知語を含む文を 213 文使用した．そのうち，Model1 対訳単語辞書を使用して未知語処理できた文が 86 文あった．これら 86 文において，未知語処理前の文（対訳単語辞書なし）と未知語処理後の文（Model1 対訳単語辞書を使用）を，人手による対比較評価を行った．人手評価結果を表 19 に示す． また．評価基準を以下に示す．

- 未知語処理○： 未知語が適切な対訳単語に翻訳され，未知語処理後の文の意味が読み取れる
- 未知語処理 ×： 未知語が不適切な対訳単語に翻訳され，未知語処理前の文の方が意味が読み取れる
- 差なし： 2 つの出力文の意味が読み取れない

表 19　翻訳実験の評価結果

未知語処理○	未知語処理 ×	差なし
17	26	43

表 19 より，未知語処理前の文の方が未知語処理後の文より翻訳精度が高かった．これより，対訳単語辞書の精度が翻訳精度の低下を招いていることが分かる．

8.4.4 評価例

翻訳実験において，未知語処理○の例を表 20 に示す．また，未知語処理 × の例を表 21 に示す．なお，下線は未知語処理を行う箇所を指している．

表 20　未知語処理○の例

入力文	その 商品 は 口コミ 宣伝 で 売れ た 。
参照文	The product sold through word-of-mouth advertising .
未知語処理前	The goods were sold by 口コミ propaganda .
未知語処理後	The goods were sold by Word-of-mouth propaganda .

表 21　未知語処理 × の例

入力文	両 法律 とも 最高 裁判所 において 無効 と さ れ た 。
参照文	Both laws were struck down by the Supreme Court .
未知語処理前	Both laws were 無効 in the Supreme Court .
未知語処理後	Both laws were operative in the Supreme Court .

表 21 のように，未知語処理が正しく行われなかった例は複数ある．この例では，使用した対訳単語辞書において，“無効” に対する訳語が “operative” という誤った対応になっている．よって，未知語を正確に翻訳できないと翻訳精度が低下する．これは，未知語処理の困難性を示していると考える．

9 おわりに

本研究では，対訳単語確率の計算において IBM Model1 〜Model5 と HMM を用いて頻度 1 単語の調査を行った．その結果，どのモデルにおいても翻訳精度に大きな差は見られなかった．この結果から，Model4 や Model5 などの複雑なモデルを使わず，Model1 を使用しても同等の翻訳精度を得られると考えている．

参考文献

[1] 川原宰, 村上仁一 “日英翻訳における IBM Model1 を用いた未知語処理”, 言語学処理学会第 24 回年次大会, 2018.
[2] 中村友哉 ，村上仁一 “対訳単語辞書の精度調査”, 言語処理学会第 23 回年次大会, 2017.
[3] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert, L.Mercer： “The mathematics of statistical machine translation:Parameter Estimation”, Computational Linguistics, Volume 19, Number 2, pp.263-311, 1993.
[4] Stephan Vogel, Hermann Ney, and Christoph Tillmann. “Hmm-based word alignment in statistical translation”, In Oroceedings of the 16th conference on Computational linguistics-Volume 2, pp.836-841. Association for Computational Linguistics, 1996.
[5] GIZA++ : http://www.fjoch.com/GIZA++
[6] 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 ”, 第一回コーパス日本語学ワークショップ, pp.119-130. 2012.
[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate. *In Proceedings of ICLR*, 2015
[8] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. “Opennmt: Open-source toolkit for neural machine translation”, 2017.