統計翻訳における単文・重文複文の翻訳精度の評価

猪澤雅史　村上仁一　徳久雅人　池原悟
鳥取大学 工学部 知能情報工学科
{s042009,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

現在，機械翻訳において，自動的に翻訳知識を構築することができる統計翻訳が注目されている．しかし，日英統計翻訳では，ATR による旅行会話タスクと特許翻訳タスクのようにドメインの違いによる翻訳精度の報告はされていても，文の構造の違いによる翻訳精度の報告は見当たらなかった．

本研究では，辞書の例文をドメインとして，単文コーパスと重文複文コーパスに分類し，それぞれの翻訳精度の評価を行なう．また，言語モデルや翻訳モデルに関する基本的な評価も行なう．

2　統計翻訳

2.1　基本モデル

日英統計翻訳は，日本語文 j が与えられたとき，全ての組み合わせの中から確率が最大になる英語語文 $\hat{e}$ を探索することによって翻訳を行なう．以下にその基本モデルを示す．

$$\begin{aligned}\hat{e} &= argmax_e P(e \mid j) \\ &= argmax_e P(j \mid e)P(e)\end{aligned}$$

$P(j \mid e)$ は翻訳モデル，$P(e)$ は言語モデルと呼ぶ．

2.2　翻訳モデル

翻訳モデルは日本語の単語列から英語の単語列へ確率的に翻訳を行なうためのモデルである．翻訳モデルはフレーズテーブルと呼ばれる表で管理される．以下にその例を示す．

庭 ||| The garden ||| 0.043 0.38 0.05 0.02
庭 から ||| from the garden ||| 0.5 0.17 1 0.03

左から，日本語フレーズ，英語フレーズ，フレーズの英日翻訳確率 $P(j \mid e)$，英日方向の単語の翻訳確率 (IBM モデル) の積，日英翻訳確率 $P(e \mid j)$，日英方向の単語の翻訳確率 (IBM モデル) の積である．また，フレーズテーブルを生成する際，フレーズ中の単語数の上限として，max-phrase-length が定義されている．例えば，max-phrase-length が 7 の場合，英日いずれかのフレーズ中の単語数が，8 以上の単語列は生成されない．

2.3　言語モデル

言語モデルは単語列に対して，それらが起こる確率を与えるモデルである．日英翻訳では，言語モデルを用いて，訳文候補の中から英語として自然な文を選出する．

言語モデルとして代表的なものに N-gram モデルがある．N-gram モデルは，“単語列 $w_1, w_2, ..., w_n$ の i 番目の単語 w_i の生起確率 $P(w_i)$ は直前の $(N-1)$ 単語に依存する”，という仮説に基づくモデルである． 音声認識では，3-gram モデルが広く用いられている．

3　実験環境

3.1　翻訳モデルの学習

翻訳モデルの学習には，多くの方法がある．本研究では，IBM model1〜5[1] に基づく，“GIZA++[2]” を利用した，“train-phrase-model.perl[3]” を用いる．

3.2　言語モデルの学習

言語モデルの学習には，“SRILM[4]” の ngram-count を用いる．なお，N-gram モデルでは，確率が 0 となるのを防ぐために，スムージングによって近似を行なう．スムージングには多くの手法があるが，本研究では，“-ndiscount” を用いる．

3.3　デコーダのパラメータ

デコーダには “moses[5]” を使用する．moses のパラメータはパラメータチューニングを行なわず，デフォルトの値を利用する．ただし，翻訳モデルには，日英翻訳確率と英日翻訳確率の共起確率を用いたほうが良い結果が得られる [6]．そこで，“weight-t” は “0.5 0.0 0.5 0.0 0.0” とする．また，日本語から英語への翻訳では，動詞の位置が大きく変化する．そこで，“distortion weight” は “0.2” とする．

3.4　学習データ

3.4.1　単文コーパス

実験には，辞書の例文から抽出した，単文コーパス 182,899 文 [7] を用いる．単文コーパスから，Open テストデータ 1,000 文をランダム抽出し，学習には 181,988 文を用いる．学習データ量と精度の関係を調べるために 181,899 文から 1,000 文，5,000 文，10,000 文，50,000 文，100,000 文をランダムに抽出し実験を行なう．日本

語と英語の単語数を表 1 に示す．

表 1: 単文コーパスの単語数

学習データ (文)	日本語	英語
1,000	9,432	8,503
5,000	48,200	43,053
10,000	91,460	80,307
50,000	497,893	44,4004
100,000	1,006,954	851,725
181,899	1,916,262	1,648,795

3.4.2 重文複文コーパス

実験には，辞書の例文から抽出した，重文複文コーパス 122,719 文 [8] を用いる．重文複文コーパスから，Open テストデータ 1,000 文をランダム抽出し，学習には 121,719 文を用いる．単文コーパスと同様に 121,719 文から 1,000 文, 5,000 文, 10,000 文, 50,000 文, 100,000 文をランダムに抽出し実験を行なう．また，重文複文と単文では単語数に違いがある．重文複文コーパスの日本語と英語の単語数を表 2 に示す．

表 2: 重文複文コーパスの単語数

学習データ (文)	日本語	英語
1,000	13,663	11,050
5,000	69,107	56,132
10,000	138,109	112,136
50,000	691,893	560,389
100,000	1,381,961	1,119,533
121,719	1,711,869	1,378,791

表 1 と表 2 を比較すると，重文複文の単語数は単文の単語数より 3 割ほど多いことがわかる．

3.5 評価方法

出力文の評価には自動評価法である BLEU[9] と METEOR[10] を使用する．BLEU は 4-gram が正しい場合に，METEOR は単語属性が正しい場合にそれぞれ高いスコアを出す．また，どちらの評価も 0 から 1 の間で評価され，1 が最も良い評価である．

4 統計翻訳の基本的な評価

4.1 max-phrase-length と翻訳精度の関係

4.1.1 目的

英仏翻訳のように似た言語間では，単語の位置に大きな変更はない．しかし，日英翻訳のように動詞の位置が大きく異なる言語間では，フレーズテーブルの単語列の長さは長いほうが、翻訳精度は高くなると考えられる．そこで，max-phrase-length と翻訳精度の関係を調べる．

4.1.2 実験

実験は，学習データに重文複文コーパス 121,719 文を用い，評価には重文複文を用いる．結果を表 3 に示す．

max-phrase-length の値を大きくすることで，翻訳精度が向上している．しかし，Open テストにおいて，max-phrase-length20 と 100 の間で精度の向上がない．

表 3: max-phrase-length と翻訳精度の関係 (重文複文，121,719 文)

max-phrase-length	Closed		Open	
	BLEU	METEOR	BLEU	METEOR
5	0.235	0.450	0.082	0.315
10	0.492	0.645	0.092	0.323
20	0.895	0.930	0.101	0.332
100	0.901	0.934	0.101	0.332

4.1.3 結論

4.1.2 節の実験から，max-phrase-length の値は 20 が適当であると考えている．以降の実験ではこの値を用いる．

4.2 *N*-gram と翻訳精度の関係

4.2.1 目的

言語モデルは，学習データが大量にある場合，次数の大きい *N*-gram モデルほど翻訳精度が高くなると考えられる．しかし，学習データ量が限られるため，次数の大きい *N*-gram モデルはパラメータの数が多くなり，信頼性が低くなる．そこで，言語モデルの *N*-gram と翻訳精度の関係を調べる．

4.2.2 実験

実験は，学習データに単文コーパス 181,899 文を用い，評価には単文を用いる．翻訳実験の結果を表 4 に示す．

表 4: *N*-gram と翻訳精度の関係 (単文，181,899 文)

言語モデルの N-gram	Closed		Open	
	BLEU	METEOR	BLEU	METEOR
1-gram	0.699	0.839	0.059	0.293
2-gram	0.789	0.856	0.113	0.366
3-gram	0.847	0.895	0.135	0.386
4-gram	0.852	0.890	0.139	0.388
5-gram	0.853	0.901	0.139	0.389
6-gram	0.853	0.901	0.139	0.388
7-gram	0.853	0.901	0.139	0.388

結果から，Closed テスト，Open テストともに，5-gram のとき翻訳精度が最も高くなっていることがわかる．また，5-gram 以上では翻訳精度が変化しない．

4.2.3 結論

4.2.2 節の実験から，日英間の翻訳では，5-gram の言語モデルを用いるのが適当であると考えている．以降の実験ではこの値を用いる．

4.3 言語モデルと翻訳精度の関係

4.3.1 目的

統計翻訳では様々な研究が行なわれているが，日英翻訳において，言語モデルの学習データ量と翻訳精度の関係の報告は見あたらなかった．そこで，翻訳モデルの学習データ量を一定にしたときの，言語モデルの学習データ量と翻訳精度の関係を調査する．

4.3.2 実験

実験は，学習データに単文コーパス 1,000〜181,899 文を用い，評価には単文を用いる．結果を表 5 に示す．なお，翻訳モデルの学習データ量は常に 181,899 文である．

表 5: 言語モデルと翻訳精度の関係
(単文，翻訳モデル，181,899 文)

言語モデルの	Closed		Open	
学習データ (文)	BLEU	METEOR	BLEU	METEOR
1,000	0.979	0.985	0.053	0.264
5,000	0.954	0.967	0.077	0.305
10,000	0.935	0.953	0.085	0.316
50,000	0.887	0.923	0.116	0.363
100,000	0.870	0.911	0.130	0.379
181,899	0.853	0.901	0.139	0.389

4.3.3 結論

翻訳モデルの学習データを一定にし，言語モデルの学習データ量を増加した場合に，学習データ量に対して，翻訳精度はほぼ線形に変化することがわかる．

4.4 翻訳モデルと翻訳精度の関係

4.4.1 目的

4.3.1 節と同様に，日英翻訳において，翻訳モデルの学習データ量と翻訳精度の関係の報告は見当たらなかった．そこで，言語モデルの学習データ量を一定にしたときの，翻訳モデルの学習データ量と翻訳精度の関係を調査する．

4.4.2 実験

実験は，学習データに単文コーパス 1,000〜181,899 文を用い，評価には単文を用いる．結果を表 6 に示す．なお，言語モデルの学習データ量は常に 181,899 文である．

表 6: 翻訳モデルと翻訳精度の関係
(単文，言語モデル，181,899 文)

翻訳モデル	Closed		Open	
学習データ (文)	BLEU	METEOR	BLEU	METEOR
1,000	0.988	0.992	0.012	0.126
5,000	0.980	0.986	0.032	0.200
10,000	0.973	0.983	0.044	0.226
50,000	0.914	0.943	0.093	0.321
100,000	0.882	0.920	0.113	0.353
181,899	0.853	0.901	0.139	0.389

この結果から，Open テストにおいて，1,000 文から 5,000 文の間で翻訳精度の向上が大きいことがわかる．

4.4.3 結論

言語モデルを一定にし，翻訳モデルの学習データ量を増加した場合，翻訳精度が大きく変化することがわかる．特に，学習データ量が少ないときに，精度の変化が著しく，4.3 節のように線形にならない．

5 単文・重文複文の学習データとしての効果

5.1 目的

日英翻訳において，タスクの違いによる翻訳精度の報告は行なわれているが，文の構造の違いによる翻訳精度の報告は見当たらない．そこで，単文，重文複文コーパス各 100,000 文を学習データに用いて，それぞれの翻訳精度を調べる．

5.2 学習データと同じ構造の文の翻訳

学習データとして単文の翻訳には単文を用いる．また，重文複文の翻訳には重文複文コーパスを用いる．翻訳実験の結果を表 7 に示す．

表 7: 学習データと同じ構造の文の翻訳
(学習，100,000 文)

評価方法	単文の翻訳	重文複文の翻訳
BLEU	0.108	0.076
METEOR	0.346	0.299

結果から，単文と比較して重文複文の翻訳精度は低いことがわかる．

5.3 学習データと異なる構造の文の翻訳

単文の翻訳に学習データとして重文複文コーパスを用いる．また，重文複文の翻訳には単文コーパスを用いる．結果を表 8 に示す．

表 8: 学習データと異なる構造の文の翻訳
(学習，100,000 文)

評価方法	単文の翻訳	重文複文の翻訳
BLEU	0.115	0.108
METEOR	0.328	0.313

表 7 と比較すると，重文複文の翻訳精度は大きく向上し，学習データとして単文を用いることが有効であることがわかる．

5.4 人手評価

表 7,8 の実験について，人手による 4 段階評価を行なう．評価基準は以下に示す．

評価 A 出力文は入力文の訳文として問題がない．

評価 B 出力文は文法的に間違っていたり，情報が欠けていたりと，不自然さを伴うが，原文で伝えたい情報が容易に理解できる．

評価 C 出力文は多くの情報が抜けているが重要な情報は含まれており，文脈や断片的な情報から原文で伝えたい情報が理解できる．

評価 D 出力文では伝えたい情報が理解できない．または，異なって理解される．

評価結果を表 9 に示す．

表 9: 人手評価

入力文	単文の翻訳		重文複文の翻訳	
学習データ	単文	重文複文	単文	重文複文
評価 A	12/100	7/100	4/100	3/100
評価 B	22/100	21/100	9/100	9/100
評価 C	10/100	21/100	18/100	14/100
評価 D	56/100	51/100	69/100	74/100

結果から，以下のことがわかる．

1 重文複文の翻訳では，学習データに単文を用いた場合のほうが，意味が理解できる訳出が多い．

2 単文の翻訳では，学習データに単文を用いた場合，重文複文を用いた場合と比較して，意味が理解できる (A,B,C) 訳出は少ないが評価 A が多い．

3 自動評価手法は人手評価と相関性があった．

5.5 考察

5.5.1 対訳データの問題点

重文複文を翻訳する場合，学習データに重文複文を用いたとき，出力文は意味が理解できる文が少なかった．この原因について，次のように考えている．対訳データにおいて，日本語文が重文複文であっても，英語文は意訳され単文となっていることが多い．そのため，正しいフレーズ対が生成されず，意味が理解できない文が多く生成された．

5.5.2 単語選択の問題点

学習データに重文複文コーパスを用いた場合では，5.4.1 節に加えて，単語選択の間違いが多い．例として，以下のような訳出があった．なお，出力文 1 は評価 A，出力文 2 は評価 D である．

入力文 大切 な 文 が 抜け て いる

正解文 The important passage is left out

出力文 1(単文) The important sentence is missing

出力文 2(重文複文) The style is important to go out

重文複文は，単文に比べて 1 文中の単語数が多い．そのため，多くのフレーズ対を生成することができる．しかし，不適切なフレーズ対も多く生成され，出力文 2 のように誤訳を含む文が出力されると考えている．

このことから，重文複文は学習データとしてそのまま使うには有効ではなく，フレーズテーブルのクリーニングを行ない，不適切なフレーズ対を削除する必要があると考えている．

5.5.3 未知語の問題

学習データに単文コーパスを用いた場合，出力文の多くは必要な単語が翻訳されず，未知語となって出力されていた．単文は 1 文中の単語数が少ないため，未知語が発生しやすい．

このことから，単文を学習データとして使うのであれば，より大規模なコーパスを用いる必要がある．もしくは，未知語対策が必要であると考えている．

6 おわりに

本研究では，単文コーパス 181,899 文，重文複文コーパス 121,719 文を学習データに用いて単文・重文複文の翻訳精度の評価を行ない，統計翻訳の基本的な評価と問題点を検討した．

実験の結果，max-phrase-length の値は 20 が適切であり，言語モデルは 5-gram を用いることが適切であることが確認できた．また，言語モデルは学習データ量に対して，翻訳精度を線形に向上させ，翻訳モデルは非線形に向上させることがわかった．最後に，重文複文の翻訳に，学習データとして単文が有効であることがわかった．

今後は，フレーズテーブルのクリーニングを行ない，その有効性を調査することを考えている．

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Matematics of Statical Machine Translation, Parameter Estimation”, Computational Linguistics, 19(2), 1993.

[2] GIZA++, http://www.fjoch.com/GIZA++

[3] training-release-1.3.tgz, http://www.statmt.org/wmt06/shared-task/baseline.html

[4] SRILM, The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm

[5] Moses, moses.2007-05-29.tgz, http://www.statmt.org/moses/

[6] Jin’ichi Murakami, Masato Tokuhisa, Satoru Ikehara, “Statistical Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables”, International Workshop on Spoken Language Translation 2007, pp.151-155, 2007.

[7] 西山七絵, 村上仁一, 徳久雅人, 池原悟,“単文文型パターン辞書の構築”, 言語処理学会第 11 回年次大会, pp.372-375, 2005.

[8] 村上仁一, 池原悟, 徳久雅人, “日本語英語の文対応の対訳データベースの作成”, 「言語，認識，表現」第 7 回年次研究会, 2002.

[9] NIST Open Machine Translation, http://www.nist.gov/speech/tests/mt

[10] The METEOR Automatic Machine Translation Evaluation System, http://www.cs.cmu.edu/ alavie/METEOR/