特定分野の統計的日英機械翻訳における学習データに関する検証
Verification of training data in Japanese-English statistical machine translation for specific domain

小林　洋也
Hiroya Kobayashi
法政大学情報科学部ディジタルメディア学科
E-mail: hiroya.kobayashi.kt@cis.hosei.ac.jp

Abstract

In machine translation, there is a technique, statistical machine translation that work machine translation by perform enormous quantity of parallel translation sentences by statistical means likes google. But there are some problems that word selection or order of word, alignment in decoding of statistical machine translation, so average translation accuracy is low yet. And so, in this paper, statistical machine translation was implemented and means that studied corpus in specified content and means that alignment in phrase to take a little heat off problems that word selection and alignment was proposed and was tested. As a result, translation that translated sentence in specified content was high score by means that studied corpus in its content. And means that alignment in phrase was felt the capability. In addition, efficient means to step up translation accuracy or the case of easy to work translation by means that changed the amount of studied data or input sentence was considered. As a result, study type of parallel translation sentences for translation accuracy improvement in statistical machine translation was effective was discovered.

1 まえがき

インターネットなどの情報技術の進歩と普及により,10年前と比べて私たちが手にすることの出来る電子化された情報の量は飛躍的に増えた.また,それらの私たちの手にすることの出来る情報には母国語によって書かれたものだけではなく,全世界のありとあらゆる言葉によって書かれたものが多く含まれている.しかし,例えば私たち日本人が英語の Web ページに辿り着いた時,残念ながら,全ての人が難なくその Web ページの内容を理解できるとは言い難い.そのため,コンピュータを利用して他言語を別の言語に翻訳する,機械翻訳の技術の需要や利用度は極めて高いのが現状である.

従来の統計的機械翻訳では辞書や文法規則などを人手で与え,それに基づいた翻訳を行う手法が用いられている[1].しかし,それは多大な時間と労力を必要とし,ある言語間で新しく機械翻訳を構築する場合などには効率的な手法ではない.そこで,労力を抑えて機械翻訳を実現する手法として,統計的機械翻訳が注目されている.統計的機械翻訳は膨大な対訳文に統計的な処理を施すことで機械翻訳を行う手法であり,自動で学習することが出来るため,従来法に比べ少ないコストで機械翻訳を行うことが期待出来る.

しかし,一般には統計的機械翻訳による翻訳精度はまだ低いのが現状である.そこで本研究では実際に自ら統計的機械翻訳を実装し,そのうえで訳語選択とアライメントの問題を解消し,翻訳精度を向上させるための実験を行う.

Supervisor: Prof. Katsunobu Ito

2 統計的機械翻訳

2.1 技術概要

統計的機械翻訳では翻訳先言語の文Ｅは翻訳元言語の文Ｊが雑音のある通信路を通ったことで変換されたものであり,翻訳を文ＪからＥへの復号化であると考える[2].統計的機械翻訳の概要図を図 1 に示し,以下で各項目を説明する.本研究では日本語から英語への翻訳を行う.

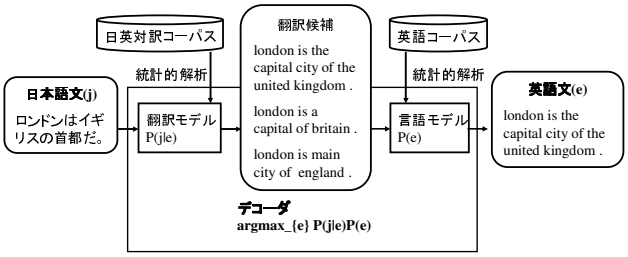

図 1:統計的機械翻訳の概要

2.2 対訳コーパス

統計的機械翻訳では同じ内容について２つの異なる言語で記述された,対訳コーパスを学習データとして用いる.本研究では日英翻訳を行うため,使用するコーパスは日英対訳コーパスを使用する.

2.3 翻訳モデル

翻訳モデル $P(j|e)$ を使って翻訳文の候補を生成する.翻訳モデルは対訳ペアごとに単語の対応をとったアライメントという概念を利用して,すべての単語対応に関する条件付確率 $P(j,a|e)$ の和として以下のように表すことができる.

$$P(f|e)=\sum_{a}P(f,a|e) \qquad (1)$$

本研究では IBM 翻訳モデル[3]を利用する.『GIZA++』を使って単語翻訳確率と単語アライメントの計算を行う.単語翻訳確率とは,ある単語が別の言語のある単語に翻訳される確率のことである.例えば”ロンドン”という単語が”london”という単語に翻訳される確率が 0.74 であるとき,「$P(london|ロンドン)=0.74$」のように書く.また,単語の対応付けを表す単語アライメントは『GIZA++』では以下のような形で表される.

ロンドン は イギリス の 首都 だ .
NULL ({ 6 }) london ({ 1 }) is ({ 2 }) the ({ }) capital ({ 5 }) city ({ }) of ({ 4 }) the ({ }) united({ 3 }) kingdom ({ }) .({ 7})

これは例えば,"london({ 1 })"は"london"という英語の単語が,日本語の文の"1"番目の位置,即ち日本語の単語"ロンドン"に接続していると見ることが出来る.

2.4 言語モデル

言語モデル $P(e)$ は単語単位のN個連続をマルコフモデルで示した N-gram 確率で計算する.例えば"london"という単語の後に"is"という単語が続く確率が 0.08 であるとき,「$P(\text{is} \mid \text{london}) = 0.08$」のように書く.

統計的機械翻訳では,言語モデルを使って英語らしい候補のスコアを高くし,翻訳モデルによって作成された候補文の中から最もスコアの高かった文を翻訳文とする.

3 デコーダ『Moses』

デコーダでは,翻訳モデルと言語モデルを利用して,最適な翻訳解 $\hat{e}$ を $\hat{e} = \arg\max_{e} P(j \mid e)P(e)$ で計算する.本研究では『Moses [5][6]』というフリーのデコーダを使用する.

『Moses』では翻訳モデルから計算された単語翻訳確率と単語アライメントを使って,フレーズテーブルを作成する.フレーズテーブルでは単語アライメントを基に推定した単語或いは単語列をフレーズと呼び,2言語間のフレーズ対に対して単語翻訳確率などを元にフレーズ翻訳確率を計算し与える.

『Moses』ではフレーズの並び替え確率を Reordering モデルと呼び,前のフレーズアライメントとの関係から 3 パターンに分け,それぞれに確率を与える.

『Moses』では与えられた入力文に対し上記のフレーズテーブルと Reordering モデルのパラメーターから図 2 のように翻訳候補を生成する.

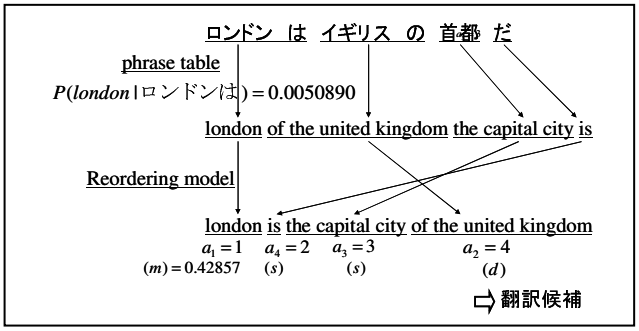

図 2:翻訳候補の生成例

4 提案手法

4.1 デコーディングの問題点

統計的機械翻訳のデコーディングの問題点としては主に以下のようなことが知られている[7][8].

・訳語選択問題
"bank"の訳語として"銀行"や"土手"といったものが存在したときに,"bank"の訳語としてどれを選択するのかという,最適化問題.

・単語並び変え問題
翻訳した単語を最も文章らしくするためにはどの単語の並びにすればよいのかという問題.NP 完全問題.

・アライメント
単語でアライメントをとると,考えられるアライメントの数が多く,最適なものが得にくいという問題.

4.2 では訳語選択を,4.3 ではアライメントの問題をそれぞれ軽減し翻訳精度を向上させるための手法について述べる.また,4.4 と 4.5 では現状の統計的機械翻訳で翻訳が行い易い場合や,今後さらに翻訳精度を向上させるために有効な手段を検証するための実験手法の提案を行う.

4.2 特定の内容による学習

統計的機械翻訳ではコーパスの内容を考えた処理は行っていないが,特定の内容のコーパスを学習データとして用いることでその内容でよく用いられるイディオムや単語を学習し,他の内容の入力文に比べ学習したコーパスの内容の入力文で高いスコアの高い翻訳文が得られるのではないかと考えた.そこで,訳語選択問題に対する提案手法として"情報科学"の内容が書かれたコーパスを学習データにし,それに対して"情報科学"の内容の他に"経済"など基の学習データに多く存在した内容と,"会話文"などの一般的に機械翻訳での翻訳が難しいとされている内容の入力文を与え,内容別の翻訳精度を検証した.

4.3 フレーズアライメント

本研究で使用しているデコーダ『Moses』では単語アライメントなどの情報を元に翻訳モデルを作成,その後のフレーズテーブルの作成に使用しているが,ここでは単語アライメントに変わるフレーズアライメントの実験を行った.例えば"機械 翻訳"と"machine translation"のアライメントを考えたとき,単語アライメントでは接続の組み合わせは6通り存在する.しかし"機械翻訳"と" machine translation"をそれぞれ,一つの単語であると見なしてしまえばその接続方法は1通りのみになり,間違いを大幅に軽減出来る.

4.4 学習データ量による翻訳精度の変化

通常,統計的機械翻訳では学習するデータ量が増えれば増えるほど翻訳精度は上がっていくとされている.そこで学習データである対訳コーパスの量とそれに伴う翻訳精度の変化について実験を行った.

4.5 入力文による翻訳精度の変化

統計的機械翻訳では一般的にどのような文章なら高い翻訳精度を誇るのかということと,入力文と学習データの類似度が翻訳精度にどう影響するのかということの検証を行った.類似度は(2)式で計算することのできる,パープレキシティという言語モデルの評価尺度を用いて計算した.

$$P = 2^{H} \qquad H = -\frac{1}{n}\sum_{i=1}^{n} \log P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \qquad (2)$$

5 使用データと評価手法

5.1 使用データ

本研究で統計的機械翻訳の実装に使用したデータを表 1 に内容別に記述する.ここで単語数とは単語の種類の数のことを意味している.

表 1:日英対訳コーパス

コーパス	文章数	日本語 単語数	英語 単語数
ロイター [9]	58,397	710,183	1,427,549
読売新聞記事[9]	14,750	201,725	492,822
The Dairy Yomiuri[9]	133,261	1,359,180	3,249,600
英辞郎	183,905	579,258	1,373,144
小説など[10]	52,327	324,057	826,258

言語モデルは,433,303 文(7,454,184 単語, のべ 41,443,519 語)の英語コーパスから SRILM[11]を用いて作成した.

5.2 BLEU

翻訳結果客観的に評価することは重要である.本研究では BLEU(Billingual Evalution Understudy)という評価手法によって翻訳精度を比較する.

BLEU は,システムの出力と参照文の間における n-gram の一致度を用いる評価手法である.フランス語などの文章を英語に翻訳するシステムの性能を,人手による評価と高い相関をもって求められることが確認されている. 辞書などの言語知識が必要ない客観性の高い手法である.[12]

BLEU では翻訳文のスコアを以下の式で計算する.

$$BLEU = BP \times \exp(\sum_{n=1}^{N}(\frac{1}{N}\log P_n)) \quad (3)$$

$$P_n = \frac{n-gramの一致数}{翻訳文の長さ} \quad (4)$$

ここで N は n-gram の最大値である.通常 N=4 が最もよいとされているが,本研究ではまだ翻訳結果があまり良くない点も考慮して,N=1 の uni-gram,すなわち単語の翻訳結果のみでスコアを計算する.

図 3 に"ロンドンはイギリスの首都である."という入力文の,BLEU スコアの計算例を示す.

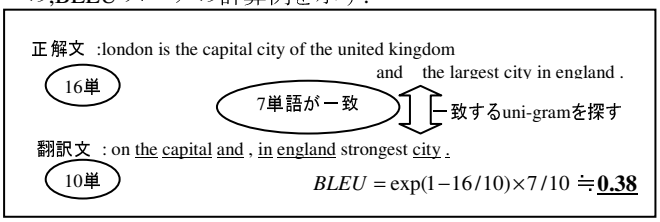

図 3:BLEU スコア計算例

6 実験結果

6.1 特定の内容による学習

学習データは表 1 の全対訳コーパスの中から"情報科学"の内容の文章 10 万文を選出して使用した. "情報科学" の文章選出はインターネットサイト『IT 用語辞典 e-Words[13]』で紹介されている IT 用語 8904 語のうち 2 つ以上を含む対訳文の抽出をプログラムで行った.学習データに含まれていた日本語単語数,英語単語数は以下の通り.

文章数 :	日本語単語数	英語単語数
100,000 :	13,960,078	17,216,440

評価用データは表 1 の全対訳コーパスの中から,学習に使用しておらず且つ"情報科学","経済","政治", "小説","会話文"の 5 つの内容の文章 50 文を人手で選出し使用した.

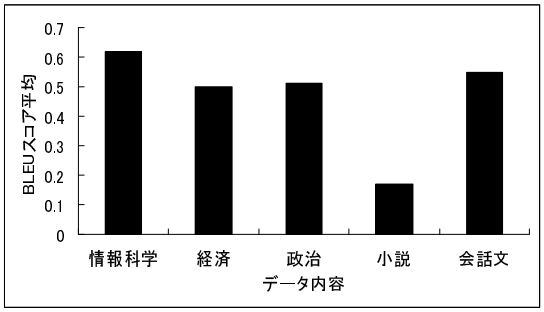

図 4:コーパスの内容と翻訳精度の変化

図 4 から翻訳結果 50 文の BLEU スコア平均は,学習したコーパスの内容と同じ"情報科学"のものが最も高かった.この結果を検証するために,「"情報科学"の内容の翻訳結果が,他の内容の翻訳結果より良いのは偶然である.」という帰無仮説に対して t 検定を行った.その結果,$t(248)=4.832$, $p<.05$ となり帰無仮説は棄却された.すなわち,この実験によって「"情報科学"の内容の翻訳結果が,他の内容の翻訳結果より良いのは偶然ではない.」という結果を得た.

6.2 フレーズアライメント

実験では全てのフレーズについてアライメントをとることは難しかったため,自らが選んだ特定のフレーズについてのみを実験対象とした.方法としてはフレーズの単語区間のスペースをとってしまうことで,フレーズを単語と認識させた.フレーズは学習に使用した対訳コーパスの中から 50 回以上出現しているフレーズ対を選出して利用した.例えば,"通貨 同盟"という日本語のフレーズと"monetary union"という英語のフレーズのフレーズ対の出現回数は 57 回であった.この他に 9 つ,合わせて 10 つのフレーズペアを選出して実験を行った.

学習データは後述する「6.3 学習データ量による翻訳精度の変化」での 10 万文の対訳コーパスを使用した.

評価用データは表 1 の全対訳コーパスの中から,学習に使用しておらず且つ①～⑩までのフレーズを含む文章を,それぞれランダムで 50 文を選出して使用した.

単語でアライメントをとった場合とフレーズでアライメントをとった場合それぞれで①～⑩のフレーズ翻訳確率を計算した.また,①～⑩のフレーズを含む文章の翻訳を行い,BLEU スコアを計算した.その結果を図 5 に示す.

フレーズ翻訳確率,BLEU スコア平均共に,ほとんどの場合でフレーズアライメントの方が高いという結果となった.しかし,BLEU スコア平均に対して t 検定を行った結果,$t(99)=0.273$, $p<0.5$ で有意となった.

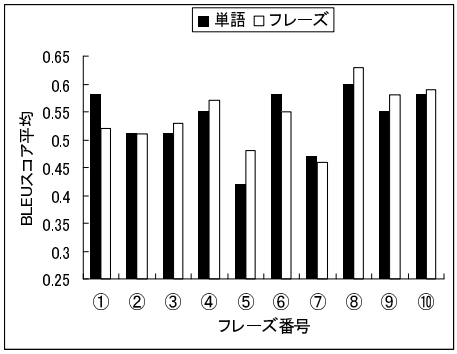

図 5:BLEU スコアの変化

6.3 学習データ量による翻訳精度の変化

対訳コーパスは 1 万から,60 万文の 13 のパターンを試す.これによって統計的機械翻訳の学習データとして対訳コーパスはどの程度必要か,学習データを増やし続けることは翻訳精度の向上にどの程度有効かを検証する.学習に使用する文章は対訳コーパス全体からランダムで選出した.

評価用データは学習に使用していない対訳コーパスの中から,ランダムで 300 文を選出し使用した.

15 万文以降は BLEU スコアの平均値が 0.3～0.35 の間で変化していることが確認出来た.また,1 万文から 10 万文ま

ではBLEUスコアの平均値が文章数に比例して上昇していることが確認できたため,1 万文ずつ 10 段階での相関係数を検定した結果,t(8)=3.03,p<0.5 で有意となった.

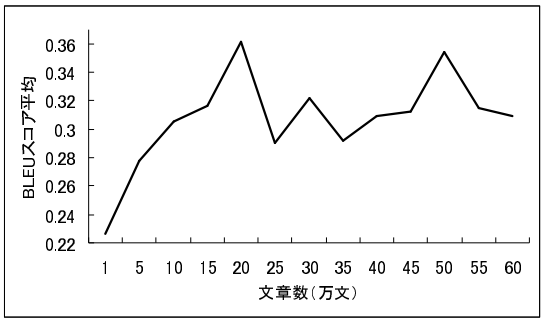

図 9: 対訳コーパス量と翻訳精度の変化

6.4 入力文による翻訳精度の変化

学習データとして,6.3 での 10 万文の対訳コーパス,評価用データとして, オープンな文章とクローズドな文章を各 250 文用意して使用した.

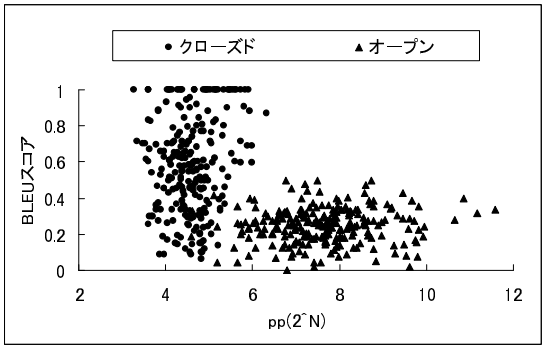

図 10:入力文と翻訳精度の変化

評価用データに使用した対訳文それぞれについて学習データとの類似度をパープレキシティ値で計算し,BLEUスコアごとにプロットした散布図を図 10 に示す.図 10 からオープンな文章よりクローズドな文章の方が,翻訳結果の BLEU スコア平均が高いという結果となった.また,BLEUスコアが高いものはパープレキシティが低いといことが分かった.しかし,オープン,クローズドそれぞれ散布は予想に反し右肩上がりとなっている.

7 考察

特定の内容による学習で“情報科学”が最も高い翻訳精度になった原因として,情報科学の単語が多く学習できたため,入力文に含まれる未知語が他と比べ圧倒的に少なかったことや,情報科学分野の単語翻訳確率が実験前と比べ上昇していたことなどが考えられる.

フレーズアライメントと単語アライメントの翻訳精度があまり変わらなかった原因として,実験したフレーズのパターンが少なく文章単位で考えた場合に依然としてアライメントのパターンが多かったことが考えられる.

学習データ量による変化では,ある学習データがある一定の数を越えると翻訳精度があまり変化しなくなる原因として,学習に使用する文章の量の他に,文章の内容や長さなどにも影響しているのではないかと考えられる.

入力文による変化では,パープレキシティの値が高い方が BLEU スコアも高いという結果になった原因として,パープレキシティの値が低くてもクローズな文章とは離れている可能性があり,類似度の尺度としてパープレキシティを用いたのが良くなかった可能性が考えられる.

8 あとがき

特定の内容による学習では,学習データの内容を限定することで,その内容に関する翻訳精度は上昇することが確かめられた.今後は訳語選択問題に関する検証のため,まずは評価用データを増やす。

フレーズアライメントでは,翻訳精度に変化は特に見られなかった.今後はフレーズパターン決定の自動化についても研究を進める.

学習データ量による変化では,ただ学習データ量を増やすだけではある程度までしか翻訳精度は向上しないということが分かった.今後は対訳文章の内容や長さなどの翻訳精度への影響についても検証を行う.

入力文による変化では,オープンな文章よりクローズド名文章の方が翻訳精度は圧倒的に良く,またパープレキシティの値が高い方がBLEUスコアも高いという結果を得た.今後は類似度を測る尺度について改めて考案をする.

文献

[1] 宮平ほか“インターネット機械翻訳の世界”毎日コミュニケーションズ

[2] 金ほか “言葉と心理の統計” 岩波書店 pp100-128(2003)

[3] Peter F. Brown, et al “The mathematics of statistical machine translation: Parameter estimation. “ Computational Linguistics, 19(2):263--311. 1993.

[4] Franz Josef Och. “GIZA++: Training of statistical translation models.” (http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.2001-01-30.tar.gz)

[5] Philippilipp Koehn, et al. Moses: a factored phrase-based beam-search decoder for machine transla. (http://www.statmt.org/moses/index.php?n=Main.HomePage)

[6] P.Koehn et al. “Statistical phrase-based translation. “ HLT-NAACL 2003 (1):48--54. 2.

[7] 渡辺ほか “階層的句アライメントを用いた統計的機械翻訳” 信学論 D-Ⅱ Vol.J-87-D-Ⅱ pp.978-986 2004.4

[8] P.F.Brown, et al. “The mathematics of statistical machine translation: Parameter estimation” Computational Linguistics, vol.19,no.2 pp.263-311, 1993

[9] M.Utiyama et al.(2003) “Reliable Measures for Alignment Japan-English News Articles and Sentence.” ACL-2003,pp72-79

[10] M.Utiyama et al.(2003) ”English-Japanese Translation Alignment Data. (http://www2.nict.go.jp/x/x161/members/mutiyama/align/download/index.html)

[11] SRI International Speech Technology and Research Laboratory. SRILM - The SRI language modeling toolkit. (http://www.speech.sri.com/projects/srilm/download.html)

[12] 金山ほか “翻訳精度手法 BLEU の日英翻訳への適用”情報処理学会 2003-NL-1

[13] IT 用語辞典 e-Words (http://e-words.jp/)