F-04

統計翻訳における未知語の対策
Measures of The Unknown Language in Statistical Machine Translation

平岡 英貴†　　山本 博史‡
Hideki Hiraoka　　Hirofumi Yamamoto

1. はじめに

近年、機械翻訳の手法として主流となりつつある統計翻訳は、翻訳元言語と翻訳先言語の対訳文を大量に集めたパラレルコーパスから、自動的にシステムを構築するものであり、専門家の知識や対訳辞書を必要としないという利点がある。統計翻訳はパラレルコーパス中の単語の出現頻度から統計量を推定し、それに基づいて翻訳するため、統計量を取る対象であるパラレルコーパスの量は多いほど望ましい。しかし、大量のパラレルコーパスを用いても出現頻度が低い単語は存在し、充分な統計量が得られないために、翻訳ができない単語が存在するという未知語の問題がある。さらに、このような未知語は文のキーワードであることが多く、未知語を翻訳できないことは文章全体の可読性低下の起因となりえる。

この未知語を減少させるため、外部の単語辞書などから対訳を得、それをパラレルコーパスに追加する手法が考えられる。しかし、この手法では、単語辞書などのパラレルコーパス以外のリソースが必要となる。

そこで本研究では、パラレルコーパスのみを利用して、未知語を削減し、文章全体の可読性を向上させることを試みる。

2. 統計翻訳

統計翻訳は、次式に示されるように与えられた翻訳元文fに対し、最も確率が高くなる翻訳先言語文eを見つける問題である。

$$\arg_e \max P(e|f) = \arg_e \max P(f|e)P(e)$$

左辺はベイズ則を用いて右辺に変換でき、この内 $P(f|e)$が翻訳モデル、$P(e)$が言語モデルと呼ばれる。言語モデルとしては、一般的に N-gram と呼ばれる直前の単語、または単語列が与えられた時、それに続く単語の生起確率を与えるものが用いられる。翻訳モデルはさらにフレーズ対訳モデルと、語順入れ替えモデルに分けられる。このうち、フレーズ対訳モデルは翻訳元言語の単語列（フレーズ）が翻訳先言語では、どのような単語列にどれくらいの確率で翻訳されるかが、テーブルの形で記憶されている。

3. 研究内容

本研究において、翻訳元言語は英語、翻訳先言語は日本語とする。

3.1 全角化処理

英日翻訳において、未知語のうちいくつかの単語は日本語においてもアルファベット表記で問題がない。このような単語は、英語における半角英数字を全角英数字に置き換えるだけで正しい翻訳となり、この処理を全角化と呼ぶことにする。しかしながら、すべての未知語を全角化すれば良いというわけではないため、全角化すべき単語なのかどうかの判定を行う必要がある。全角化するべき単語は、その文字列並びに特徴があると考えられ、その特徴を統計量でとらえることにする。手順は以下の通りである。

1. 学習データの中から未知語となりやすい低頻出単語を集め、全角化対象になるもの、すなわち対訳がアルファベットのままのものと、そうでないものの２種類に分ける。
2. その２種類に分けた単語を用いて、文字を単位とする言語モデルをそれぞれ作成する。
3. 翻訳対象文中に現れた未知語に対し、双方の言語モデルに対するクロスパープレキシティ（以下 ppl）を求める。
4. 3 で算出された ppl が、全角化対象の単語から得られた言語モデルの方が低い場合、その未知語を全角化対象単語と見なし、英語半角単語と日本語全角単語のペアをフレーズ対訳モデルに追加した後、翻訳対象文の翻訳を行う。
5. そうでない場合は、後述する別の方法を用いる。

3.2 ハイフン処理

全角化処理対象以外の単語として、いくつかの英単語がハイフンによって連結されているために翻訳できない場合がある。この場合、翻訳対象文中のハイフンをスペースに置き換えることである程度意味の通る翻訳が可能となると考えられる。そこで、翻訳対象文中のハイフンを含む未知語中のハイフンをスペースに変換した後翻訳することとする。

3.3 先頭の大文字を小文字に変換する処理

英文の場合、先頭単語の先頭文字は大文字になる。ただし、固有名詞等、元々先頭が大文字の単語もあるため、単純に小文字に戻すだけではすまない。そこで、翻訳対象文中の未知語のうち、大文字で始まる英単語を、小文字で始まる英単語に変換してから翻訳を行う。

3.4 複数形と過去形処理

その他の未知語としては、複数形英単語、過去形英単語が存在する。この場合，それぞれの単数形、現在形は未知語でない場合が考えられる。

日本語では、単複を区別しないため、複数形の英単語は単数形の英単語に変換した後翻訳しても問題がない場合が多いと考えられる。よって、翻訳対象文中の複数形の未知語を単数形に置き換える処理を行う。

また、過去形英単語は、単に現在形英単語に変換してしまうと、日本語訳もまた現在形になってしまう。それを避

† 近畿大学大学院生総合理工学研究科, Graduate School of Science and Engineering. Kinki University
‡近畿大学理工学部, Department of Science and Engineering. Kinki University

けるために、過去形英単語の未知語を現在形英単語に変換し、それの日本語訳を得る。そして、過去形の英単語と現在形の日本語訳を過去形に変換したものをフレーズ対訳モデルに追加した後翻訳することとする。

4. 実験と結果

以下に実験と結果を示す。

4.1 実験条件

実験に用いたデータは英日特許文約 180 万文対である。このうち 160 万分を学習用とし、残り 20 万文のうちの未知語を含む 1160 文を評価用として使用した。この中には、未知語がのべ 1315 語含まれている。ここで特許文を使用した理由として、特許文には専門用語といった未知語になりやすい単語や記号列を数多く含み、本研究における全角化処理をはじめとする未知語への対策に適していると考えられるためである。

翻訳方向は 3 章で述べたように、英語から日本語の英日翻訳であり、翻訳モデルの学習には moses[1]、言語モデルである N-gram の学習には srilm[2]を用いた。評価尺度には BLEU[3]を用いた。

4.2 全角化処理の結果

翻訳文に含まれる未知語に対して、全角に変換するべき単語と、そうでない単語を取り出し、それぞれに対し正しく判定できているかを 3.1.節に示した方法で判定を行い、その判定精度を評価した。その結果を表 1 に示す。

表 1.ppl 値による振り分けの結果と評価

	全角化と判定	非全角化と判定
全角化単語	83.33%	16.67%
非全角化単語	22.23%	77.77%

表 1 に示されるように、全角化するべき単語は約 83.33%、全角化するべき単語は約 77.77%正しく判断することができた。この結果を反映させ再度翻訳したところ、全角化を行わなかった BLEU 値が 19.50 だったのに対し、処理後は 19.63 に上昇したことが確認できた。また、語数では評価用文中の未知語 1315 語中、406 語が正しく翻訳できた。

この際、全角化、非全角化対象単語のうち、正しく判定されたものの、および誤って全角化対象単語ではないと判定された未知語の例を表 2 に示す。

表 2.判定された未知語例

	全角化と判定	非全角化と判定
全角化単語	IHARA, RAM23	5A-5B-3
非全角化単語	TAKEOUT	methodologically

表 2 より、全角化単語において、固有名詞である人物名や名称においては正しく判定されているが、先頭が数字から始まる単語は正しく判定されない場合が多かった。また、非全角化単語においては、単語全てが大文字の場合正しく判定されないことがあった。

4.3 その他の処理の結果

引き続いて、3.2 節以降で述べた全角化以外の処理を行った。その結果を表 3 に示す。

表 3.各処理による翻訳可能単語数

	翻訳できた語数
全角英単語判定処理	406
ハイフン処理	137
先頭の大文字を小文字に	20
複数形と過去形処理	28

この時点で評価用文中に含まれる全角化対象単語以外の未知語数は 909 語であった。全角化対象単語以外の処理対象となった単語を次に示す。

- ハイフン処理 ….『Electron-Beam』
- 先頭の大文字を小文字に ….『Hemisphere』
- 複数形と過去形処理 ….『restricted』

各処理を行った所、この内の 185 語が翻訳可能となった。再度翻訳したところ、全角化処理後ほど BLEU 値の変化は見られなかったが、未知語の数は減少し、可読性が向上した。全角化とハイフン処理の翻訳例を次に示す。

- ◆ FIG.2 is a circuit diagram showing the construction of a conventional electric discharge machining apparatus of the capacitor–switching type .
- ◆ 図２は、回路の構成を示す図では 、従来の放電加工装置 は、コンデンサ切換方式である。

- ◆ The apparatus of the present invention preferably comprises 2B3T (two beds three towers) , like the conventional apparatus .
- ◆ また、本装置は２Ｂ３Ｔ（３）の床塔に、従来例の装置である。

5. 結論と課題

本研究では、未知語に対して外部の単語辞書等を用いることなく、統計翻訳における未知語の翻訳を可能とし、文全体の可読性の向上を試みた。

しかし、今回検討した方法には、未知語の対訳語を自動生成し、それをフレーズ対訳モデルに追加する形で翻訳しようとしている。しかしながら、語順入れ替えに関しては考慮を行っていないため、未知語の語順が正確でない可能性がある。そこで今後の課題として、未知語の翻訳処理と同時に語順等の制約を追加することで可読性はさらに向上すると考えられる。

参考文献

[1]. Moses, http://www.statmt.org/

[2]. Srilm-SRI Speech Technology and Research Laboratory, http://www.speech.sri.com

[3]. 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.