Vol.2017-NL-234 No.21
2017/12/21

日英中機械翻訳における自動評価の研究

月出絵里香[†1]　高岡詠子[†2]

概要：2020 年に予期される訪日外国人の増加に向けて，本研究室では医療現場で使用する多言語対応情報提供システムの開発に取り組んでおり，より幅広い状況に対応できるようにしたいと考えている．そのために，医療用コーパスを作成し，他のコーパスと共に評価し，比較，改善に努めている．本研究では，統計的機械翻訳において，日本語，英語，中国語の複数のコーパスの自動評価を行った．最初の実験では，一般的な英日機械翻訳モデルを用いて，医療用コーパスの評価を行った．次の実験では，基本的な日本語，英語，中国語のコーパスで作成した機械翻訳モデルを用いて，複数のコーパスの評価を行った．また，語順の制限の有無を加味し，評価を行った．最後に，後編集を行い，機械翻訳の改善を図った．発表時には，本研究室で作成した医療用コーパスの評価値も紹介できると思われる．

キーワード：機械翻訳，自動評価，医療コーパス

A Study on Automatic Evaluation of Japanese, English, and Chinese Machine Translation

ERIKA HITACHI[†1]　EIKO TAKAOKA[†2]

Keywords: Machine Translation, Automatic Evaluation, Medical Corpus.

1. はじめに

海外旅行中に体調を崩すことは決して珍しくない．当然そのような状況下では，病院の受け入れ体制や言語で困った人もいるだろう．来日する外国人も同様の経験があると予想される．外国人旅行者 163 人を対象に，日本人医療スタッフとコミュニケーションへの不安を英語や中国語，韓国語を用いたアンケートの調査結果によると，英語で解答した人の内 51%，中国語や韓国語で解答した人の内 57%，が不安を感じていることが示唆された[1]．従って，調査結果で示されたように多くの外国人旅行者が医療現場におけるコミュニケーションに不安を抱いている可能性が考えられる．しかし，不慣れな土地で外国人が受診することを想定している病院を探すのは容易ではないだろう．外国人が気軽に使える情報提供をするシステムがあれば，来日した際の外国人の不安要素を 1 つ取り除くことができると考えられる．

本研究室では，来日する外国人旅行者が大幅に増加すると予想される 2020 年の東京オリンピック・パラリンピック開催に合わせて，「医療・看護・福祉・介護分野における多言語対応情報提供システム」の開発に取り組んでいる．実装されれば，病院の案内の情報をまとめたり，日本語で書かれた問診票を外国人の母国語に変換したり，医療スタッフの案内する際の声かけを外国語で紹介したり，と使用用途は多岐にわたる．このような医療情報をタブレット対象のアプリを通して計 10 カ国語で提供するシステムの開発に取り組んでいる．日本語の他に英語，北京語，ミャンマー語，フランス語，タイ語，ロシア語，韓国語，スペイン語，ポルトガル語，インドネシア語に対応する予定である[2]．

2. 多言語対応情報提供システム

「医療・看護・福祉・介護分野における多言語対応情報提供システム（Sophia Cross-lingual Health Assistant System，通称 SoCHAS）」とは，第 5 回（2014～15 年度）教職協働・職員協働イノベーション研究として 「2020 年東京オリンピック･パラリンピック競技大会で上智大学ができること：医療・看護・福祉・介護分野における多言語対応情報提供システム構築を目指した人的・組織的ネットワークの構築とシステムの概念設計」から始まった研究である．システム構築の最終的な目的は，日本における医師や看護師，介護福祉士，社会福祉士などの医療従事者や専門性のあるボランティアが，医療現場にて，外国人とのコミュニケーションを円滑に取るための支援をすることである．その一環として，本研究室では医療に関する情報を 10 カ国語で提供するアプリを開発している．また，本プロジェクトを推

†1 上智大学大学院理工学研究科理工学専攻
Graduate School of Science and Technology, Sophia University
†2 上智大学理工学部情報理工学科
Dept. of Information and Communication Sciences, Faculty of Science and Technology, Sophia University

進するためのコンソシアムを 2017 年 5 月に立ち上げている[3]．より幅広い状況に対応できるようにしたいと考えている．そのために，医療用コーパスを作成し，他のコーパスと共に評価し，比較，改善に努めている．作成したコーパスの評価を行う際に我々のシステムに適した評価方法は何かということを模索するために本研究を始めた．

3. 統計的機械翻訳

本研究におけるコーパス評価に用いるアルゴリズムとして統計的機械翻訳を採用している．統計的機械翻訳とは，言語モデルと翻訳モデルの 2 つを用いた機械翻訳である．図 1（統計的機械翻訳[4]の図 1.より転載）は統計的機械翻訳の概要を表したものである．言語モデルとは，文法に相当する出力言語の単語の並びの尤もらしさを規定する統計モデルで，翻訳モデルとは，翻訳ルールや対訳辞書に相当する訳語の尤もらしさを規定する統計モデルである．入力言語と出力言語の文章が対となっているものをコーパスと呼び，言語モデルは出力言語の，翻訳モデルは入力言語と出力言語の両方のコーパスから学習する．本研究では無料で提供されているオープンソースの統計的機械翻訳ソフト「moses」[5]を用いることとした．

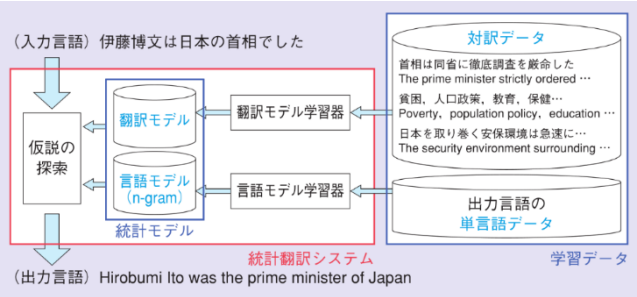

図 1　統計的機械翻訳

「moses」は様々なパラメータを自由に設定することができ，その 1 つに語順の制限というものがある．語順の制限とは，入力文を出力文に変換する際にどれくらいまで語順を移動しても良いかという数値であり，デフォルトでは 6 ワードに設定されている．欧米言語間での機械翻訳の場合，構文が似ているため語順の制限は解除しない方がいいが，英語と日本語のように構文が全く異なる場合，語順の制限を設けない方が機械翻訳の質が上がる．例えば，述語の位置が，英語では文章の始めにある主語のすぐ後に来るのに対し，日本語では述語はほとんど文章の終わりにある．これは文章が長ければ長いほど語順はだいぶ変わることになる．そのため，語順の制限は解除すれば英語と日本語の場合，機械翻訳の質が良くなると考えられる[6][7]．

4. 自動評価

本研究では，BLEU score という自動評価方法を採用している．BLEU (BiLingual Evaluation Understudy) score とは，比較する 2 つの文章の単語の並びの類似度を 0 から 100 で数値化したもので，値が大きいほど比較基の文章と並び順が似ていることを示す．一般的に BLEU score が高いほど，精度が高いと言われ，機械翻訳の精度を調べる 1 つの目安として知られている．

$$\text{bleu} = \text{BP} \times \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right) \quad (1)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c \geq r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c < r \end{cases} \quad (2)$$

図 2　BLEU score の計算式

BLEU score は図 2 の(1)の式で求めることができる．P_n とは，翻訳結果と日本語コーパスの n-gram の重なり具合を示す n-gram 精度のことである．BP とは，Brevity Penalty の略で，比較対象の文章が比較基の文章より短い場合に生じるペナルティのことを指し，図 2 の(2)の式で求めることができる．比較対象の文章が比較基の文章より長いと，BP は 1 となり，ペナルティは無いが，短いと BP は 1 未満となり，BLEU score は下がる[8]．

5. 実験

本研究では，実験を 3 つに分けて行った．

5.1 実験 1：英日翻訳の研究

1 つ目の実験では，英日翻訳のみに焦点を絞り，医療用語をコーパスに加えることで生じる評価値と翻訳結果の変化に着目した．

5.1.1 条件

この実験では，4 種類のコーパスを使用した．

・中学生の英語の教科書を基に作成されている「tanaka」(149918)[9]

・医療従事者の話などを参考にして病院での会話を中心に作成した「talk」(357)

・病院からご提供いただいた複数の問診や検査についての文書を基に作成した「hospital」(346)

・tanaka と talk を組み合わせた「tanakaPtalk」(150276)

括弧の中はそれぞれのコーパスが含む対訳の数を表している．

5.1.2 手法

tanaka を用いて tanaka モデルを，tanakaPtalk を用いて tanakaPtalk モデルを作成する．その後，tanaka モデルと tanakaPtalk モデルの両方で tanaka, talk, hospital, tanakaPtalk 全ての英語コーパスを英日翻訳した後，BLEU score を算出し，その数値を比較した．

5.2 実験 2：日本語，英語，中国語の比較，語順の制限の有無

2 つ目の実験では，日本語，英語，中国語の 3 か国語のコーパスを用いて，日英，英日，日中，中日，英中，中英の 6 種類の機械翻訳を評価した．また，語順の制限の有無による評価値の変化に着目した．

5.2.1 条件

この実験では，2 種類のコーパスを使用した．

・日本語の基本的な文を中国語と英語に翻訳した「basic」(5304)[10]

・特許の文書から集めた「ntc9」(2000) [a] [11]

5.2.2 手法

それぞれのコーパスでモデルを作成し，そのまま翻訳した後，語順の制限を解除し，もう一度翻訳した．BLEU score を算出し，その数値を比較した．

5.3 実験 3：後編集の研究

3 つ目の実験では，実験 2 の翻訳結果（翻訳結果 1 とする）を用いて，後編集を行った際の評価値を比較した．後編集とは，機械翻訳で出力された翻訳結果を修正することを指し，本研究では，翻訳結果 1 と出力言語のコーパスでもうひとつモデルを作成し，翻訳結果 1 をもう一度翻訳すること（この翻訳結果を翻訳結果 2 とする）とした．図 3 は本研究における後編集の概要を示したものである．モデルの下にはモデルの作成に使用したコーパスを記載した．後編集を行うことで翻訳結果 2 は翻訳結果 1 よりもコーパスすなわち人による翻訳結果に近づき，精度が上がると考えられる[12]．

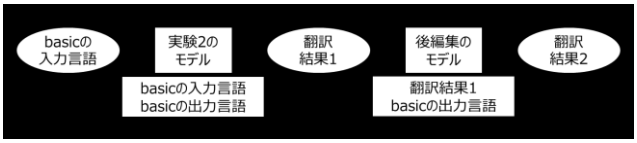

図 3　後編集

5.3.1 条件

この実験では，実験 2 と同じコーパスを使用した．

・日本語の基本的な文が集められた「basic」(5304)[7]

5.3.2 手法

実験 2 で basic を用いて語順の制限を解除したモデルで翻訳した翻訳結果を翻訳結果 1 とし、翻訳結果 1 を用いてモデルを作成した．このモデルを使用して，翻訳し，翻訳結果 2 を得た．また，実験 2 同様，語順の制限を解除し，もう一度翻訳した．BLEU score を算出し，その数値を比較した．

a) NTCIR-9 Patent MT http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-en-PatentMT.html

6. 結果と考察

6.1 実験 1

表 1　実験 1 の結果

コーパス	tanaka モデル	tanakaPtalk モデル
tanaka	48.81	46.77
talk	4.18	5.38
hospital	7.14	3.02
tanakaPtalk	48.7	46.66

実験 1 で算出した BLEU score をまとめたものが表 1 である．実験 1 では，tanaka モデルと tanakaPtalk モデルは共通して，tanaka と tanakaPtalk の値が高く，talk と hospital の値が低くなった．talk は病院での会話を基に作成したため，コーパスの大部分が口語となっていたことや tanaka には含まれていない talk や hospital で使われていた医療の専門用語が訳せていなかったことが原因で低くなったと考えられる．hospital は文書に表記されている病名などが訳せなかったことと単語になっている部分が多かったが為にうまくいかなかったことなどが原因として考えられる．

tanaka モデルと tanakaPtalk モデルの BLEU score を比較すると，tanaka，hospital，tanakaPtalk は下がり，talk は上がった．これは talk の文章を入れたことで，tanaka にしてみれば異物が，talk にしてみれば自身が，モデルの要素として組み込まれたからと考えられる．hospital は病院の文書を基に作成したため，一部訳が変わってしまったのが大きな影響を与えたと考えられる．例えば，name の訳が“名前”から“名”に変わっていた．また，talk で取り扱っている病気と hospital に載っている病気が一致していなかったため，病名の部分は tanaka モデル同様に tanakaPtalk モデルでも訳せず，医療系の文章をモデルに組み込んだにもわらず BLEU score の上昇にはつながらなかった．

表 2 実験 1 の訳の変化の例

基となる対訳	tanaka モデル	tanakaPtalk モデル
この病院は初めてですか？	は初めてですか．この	初めてですかこの
Is this your first visit to this hospital?	hospital?	hospital?.
健康保険証をお持ちですか？	ですか．あなたの健康保険カードとね．	あなたが健康保険たのでしょう．
Do you have your health insurance card with you?		
お大事に	どうぞお体を大切に．	体をお大事に．
Please take good care of yourself.		
ベッドに横になってください．	その子供はベッドに横になってください．	ベッドに横になってください．
Please lie on the bed.		

tanaka モデルと tanakaPtalk モデルの翻訳結果で変化した一例を表 2 にまとめた．翻訳結果の変化は tanakaPtalk モデルの方がより日本語らしい文章になったと考えられる．こういった変化は BLEU score には反映されず，一概に BLEU score だけでは質が判定できないことを示唆している．

6.2 実験 2

表 3 実験 2 の basic の結果

	語順の制限あり	語順の制限なし	
英日	37.67	40.25	+2.58
日英	26.05	28.5	+2.45
中日	64.97	66.85	+1.88
日中	53.92	56.83	+2.91
中英	51.01	51.58	+0.57
英中	51.28	52.09	+0.81

実験 2 の basic の BLEU score をまとめたものが表 3 である．実験 2 の basic では，中日の値が一番高く，日英の値が一番低くなった．日英，英日は全体的に低くなりやすい傾向があるため，本実験でも低くなったと考えられる．語順の制限の有無による BLEU score の変化は日中，英日，日英が高く，英中，中英が低くなった．英語と中国語は語順が似ているため，BLEU score があまり上がらず，逆に日本語と英語，日本語と中国語は語順があまり似ていないため，BLEU score が上がったと考えられる．しかし，中日は語順の制限がある状態でも BLEU score が高かった，つまり精度が高かったので，語順の制限を解除しても日中ほど上がらなかったと考えられる．

表 4 実験 2 の ntc9 の結果

	語順の制限あり	語順の制限なし	
英中	50.07	55.43	+5.36
中英	45.28	49.81	+4.53

実験 2 の ntc9 の BLEU score をまとめたものが表 4 である．実験 2 の ntc9 では，英中の方が中英よりも BLEU score が高く，語順の制限の有無の変化は約 5 上昇した．英語と中国語は語順が似ているため，語順の制限を解除しても，BLEU score は上がりにくいが，ntc9 は文章量が basic と比べ，少なく，BLEU score が低かったため，ntc9 の方が BLEU score が上昇した可能性がある．

実験 2 の basic と ntc9 両方の語順の制限の解除をする前の結果を見ると，実験 2 では，モデルに使用したコーパスを訳したため，basic は 5304 文，ntc9 は 2000 文と文章量は少ないにも関わらず，BLEU score が全体的に高くなったと考えられる． また，語順の制限を解除すると，語順の違いや言語を問わず，多少の差はあるけれども，BLEU score が上昇することが確認できた．

表 5 実験 2 の訳の変化の例

基の英文	A magnitude 1 earthquake was observed in Miyagi Prefecture.
語順の制限あり	で 震度 1 を 観測 する 宮城 県
語順の制限なし	宮城 県 で 震度 1 を 観測 する
基の日本文	宮城 県 で 震度 1 を 観測 する

basic における語順の制限の解除による翻訳結果の変化の一例をまとめたものが表 5 である．日本語と英語では，語順が全く異なるため，語順の制限を解除することで基の日本文と全く同じ結果となった．基の英文を並べ替えずに直訳すると，“震度 1 地震 観測 する で 宮城 県”となり，日本文と語順が全く違うことがわかる．語順の制限がある状態では，述語の“観測 する”が最後に来ていなかったり，“宮城 県”が文末に来ていたりと単語は訳せていても語順が異なるため，BLEU score は低かったと考えられる．

6.3 実験 3

表 6　実験 3 の結果

		後編集 (語順の制限あり)		後編集 (語順の制限なし)	
日英	28.5	46.3	+17.89	47.03	+18.53
英日	40.25	61.14	+20.89	61.58	+21.33
日中	56.83	82.01	+25.18	82.42	+25.6
中日	66.85	81.99	+15.14	82.22	+15.37
英中	52.09	73.64	+21.55	73.91	+21.82
中英	51.58	65.41	+13.83	65.67	+14.09

実験 3 の BLEU score をまとめたものが表 6 である．実験 2 の basic の結果と比較し，語順の制限がありで平均 19.08，語順の制限なしで平均 19.47 上昇した．語順の制限の解除に比べ，後編集の方が大幅に BLEU score が上昇した．

しかし，本実験では，後編集のモデルを翻訳結果ごとに作成したため，カスタマイズされている状態のため，この後編集のモデルを異なるモデルの同じ言語の組み合わせの翻訳結果に適用しても効果を発揮しない可能性もあるため，今後も検証を重ねていく必要がある．

表 7　実験 3 の訳の変化の例

基の英文	People at the company lament that ~.
実験 2/翻訳結果 1 (語順の制限なし)	の 人 が lament が ~ と ます
翻訳結果 2 (語順の制限あり)	会社 の 人 が 嘆い て ~ と ます
翻訳結果 2 (語順の制限なし)	~ と 会社 の 人 が 嘆い て い ます
基の日本文	~ と 会社 の 人 が 嘆い て い ます

実験 3 の後編集による翻訳結果の変化の一例を表 7 にまとめた．後編集によって，実験 2 の段階では訳せていなかった”lament”が“嘆いて”に変わり，足りなかった“会社”が追加され，より基の日本文に近づいた．さらに，後編集でも語順の制限を解除することで，基の日本文と全く同じ結果となった．このように，BLEU score だけでなく，翻訳結果においても大幅な改善が見られた．

7. おわりに

本研究では，3 つの実験を行い，moses を用いた統計的機械翻訳の精度向上を行った．1 つ目の実験では，英日翻訳において，医療用語をコーパスに追加することで，BLEU score と訳に変化が出た．BLEU score にはあまり変化がなかったが，訳はより日本語らしくなった．2 つ目の実験では，日本語，英語，中国語のコーパスを用いて，語順の制限の解除し，BLEU score に改善が見られた．3 つ目の実験では，後編集を語順の制限を解除して行うことで BLEU score と訳の両方に大幅な改善が見られた．これらの実験結果を踏まえて，今後は大規模なコーパスを用いて，さらに実験を重ね，より良い医療用コーパスの作成に尽力していきたいと考えている．

参考文献

[1] 山岸祥子，佐久間夕美子，宮内清子，松本彩子，堀川沙織，渋井優，青木早織，佐藤千史，外国人旅行者の我が国の医療体制に対する不安要因，Journal of International Health，Vol.23，No.4，pp273-279，(2008).
[2] 高岡詠子，医療・看護・福祉・介護分野における「多言語対応情報提供システム」，『ダイバーシティ・セミナー』＆『社会課題に応える女性研究者によるｼｰｽﾞ発表会』，2017.
[3] 高岡詠子，“SoCHAS 推進コンソシアム SoCHAS Acceleration Consortium”，https://sochas.ac.jp，（参照　2017-11-17）.
[4] 塚田元，渡辺太郎，鈴木潤，永田昌明，磯崎秀樹，統計的機械翻訳，NTT 技術ジャーナル，Vol.19，No.6，pp.23-25，(2007).
[5] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, 2007, http://www.statmt.org/moses/ (accessed 2017-11-14).
[6] 磯崎秀樹，英日翻訳における語順について，言語処理学会第 16 回年次大会　発表論文集，pp.882-887，(2010).
[7] まあ，“ようこそ統計的機械翻訳の世界へ”，Moses 奮闘記 謎多き統計的機械翻訳の世界，http://mahlog.typepad.jp/moses/，（参照　2017-11-14）.
[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp.311-318, (2002).
[9] Jim Breen, “Tanaka Corpus”, EDRDG, http://www.edrdg.org/wiki/index.php/Tanaka_Corpus, (accessed 2017-11-17).
[10] 黒橋禎夫，河原大輔，“日英中基本文データ”，黒橋・河原研究室，http://nlp.ist.i.kyoto-u.ac.jp/index.php?日英中基本文データ，（参照　2017-11-14）.
[11] Isao Goto, Bin Liu, Ka Po Chow, Eiichiro Sumita, Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. Proceeding of NTCIR-9 Workshop Meeting, December 6-9, 2011, Tokyo, Japan, pp.559-578, (2011).
[12] 園尾聡，木下聡，統計的後編集による英日・中日・韓日特許翻訳の精度向上，Japio YEAR BOOK 2015，pp.342-345，(2015).