Workshop on Asian Translation の評価結果の分析

Analysis of the Evaluation Results of Workshop on Asian Translation

京都大学大学院情報学研究科
中澤　敏明

2010 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

nakazawa@nlp.ist.i.kyoto-u.ac.jp　TEL 075-753-5346

京都大学大学院情報学研究科教授
黒橋　禎夫

1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006 年 4 月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事。

TEL 075-753-5346

1 はじめに

Workshop on Asian Translation（WAT）では毎年アジア言語を中心とした機械翻訳の評価を行なっている。2017 年のワークショップ WAT2017[1] は国際会議 IJCNLP2017 のワークショップとして台北で開催され、世界中から 12 のチームが評価タスクに参加した。評価方法は BLEU などの自動評価はもちろんのこと、人手でも行われた。人手評価はベースラインシステムとの一対評価（Pairwise Evaluation）と、特許庁が提案している「特許文献機械翻訳の品質評価手順」のうち「内容の伝達レベルの評価」[2]（JPO Adequacy Evaluation）の 2 種類の方法で行われた。JPO Adequacy Evaluation は、テストセットのうちの 200 文を対象に、2 名の評価者が以下の基準での絶対評価を行う。

本稿では WAT2017 の日英・英日特許翻訳タスクの JPO Adequacy Evaluation の結果について、特に以下の 2 つの観点で分析を行ったので報告する：

- 成績トップのシステムの評価者間一致度（κ 係数）が小さい理由
- 2 名の評価者間で評価が割れている文の検討

表 1　JPO Adequacy Evaluation の評価基準

5	すべての重要情報が正確に伝達されている。（100%）
4	ほとんどの重要情報は正確に伝達されている。（80%～）
3	半分以上の重要情報は正確に伝達されている。（50%～）
2	いくつかの重要情報は正確に伝達されている。（20%～）
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。（～20%）

2 成績トップのシステムの κ 係数が小さい理由

英日翻訳の評価結果上位 3 システムは順に SYSTEM A（平均 4.75）、SYSTEM B（4.63）、SYSTEM C（4.40）であったが、評価者間の一致度を示す κ 係数（Cohen's Kappa）は順に 0.32、0.42、0.43 であり、トップの SYSTEM A に対する κ 係数が最も小さい値となった（WAT の方針に習って、システム名を匿名化している）。これは日英翻訳においても同様で、トップのシステムの κ 係数が最も小さくなった。この理由を、英日翻訳の SYSTEM A と SYSTEM B を例にとり説明する。

表 2 に SYSTEM A の評価結果の詳細を、表 3 に SYSTEM B の評価結果の詳細を示す。

表 2　SYSTEM A の評価結果詳細

評価	5	4	3	2	1	計
5	154	11	2	1	0	168
4	11	3	3	0	0	17
3	4	3	4	0	0	11
2	0	0	2	1	0	3
1	0	0	0	1	0	1
計	169	17	11	3	0	200

表 3　SYSTEM B の評価結果詳細

評価	5	4	3	2	1	計
5	141	15	1	0	0	157
4	8	7	6	1	0	22
3	4	2	6	2	0	14
2	0	0	4	1	1	6
1	0	0	0	0	1	1
計	153	24	17	4	2	200

この表から純粋に二人の評価者の評価が一致しているものの割合 p_o を計算すると、SYSTEM A では対角線上の数字を足した 162（=154+3+4+1+0）を 200 で割って 0.81 であり、SYSTEM B では 0.78 となるため、SYSTEM A の方が一致していることになる。一方で κ 係数は二人の評価者の評価が偶然に一致する可能性も考慮している。偶然に一致する可能性 p_e は、評価者 X が評価 N をつけた割合を p_{XN} とすると、以下の式で計算される（2 名の評価者を A と B とする）。

$$p_e = \sum_{N=1}^{5} p_{AN} \times p_{BN}$$

この式に則って SYSTEM A と SYSTEM B の p_e を計算すると、それぞれ 0.72 と 0.62 となり、SYSTEM A の方が偶然に一致する可能性が高いことになる。κ 係数は上記 p_o と p_e を使って、以下のように計算される。

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

ざっくり言うと、偶然に一致する可能性を差し引いて一致率を計算していることになり、偶然に一致する可能性が高い場合ほど、κ 係数は小さく見積もられることになる。上記の場合、2 つのシステムにおいて p_o の値はさほど変わらないが、p_e の値に大きな違いがあるため、SYSTEM A の方が κ 係数が小さくなっているのである。

SYSTEM A の p_e の値が大きくなる要因の一つは、SYSTEM A の翻訳結果が良く、どちらの評価者も 5 をつける割合が高くなっているからだと考えられる。このように、非常に良いシステム（もしくは非常に悪いシステム）においては、κ 係数が不当に低く見積もられる可能性があることに注意が必要である。一般的に、評価結果がなんらかの値に偏る傾向が強い場合には、κ 係数の解釈には注意が必要であると考えられる。つまり κ 係数が低いからと言って、評価結果がばらついていると短絡的に考えることは危険なのである。

3　2 名の評価者間で評価が割れている文の検討

次に日英・英日の特許翻訳結果において、2 名の評価者間で評価が 2 以上離れている文について、どのような傾向があるのかを検討した。

3.1　評価者のミス

単純に評価者の評価ミスと思われるものがいくつか見受けられた。以下に例を示す（各翻訳文の後ろにある 1 組の数字が評価値である）。

入力	Accordingly, the present invention relates to a rod-like crystal of CZTS.
正解	それ故、本発明は、CZTS のロッド状結晶体に関する。
出力	したがって、本発明は、〈unk〉の棒状結晶に関する。(3, 5)

この例では NMT 特有の〈unk〉がそのまま出力されているにも関わらず、一方の評価者が評価 5 としている。

入力	However no long text and even only pictures may be included in pages for many book images.
正解	しかし、多くの本画像のページには、長いテキストは含まれず、ピクチャしか含まれないことがある。
出力	しかし、多くの書籍画像のページには、長いテキストや絵も含まれなくてもよい。(2, 5)

この例は入力の英語文の解釈が難しいが、正解と比較すると出力は明らかに誤っている。しかしながら一方の

評価者が評価 5 としている。

入力	A plug 72 provided on the rear end of the ink pack 7 faces the plug opening 335 .
正解	口栓用開口部 335 には、インクパック 7 の後端に設けられた口栓 72 が臨んでいる。
出力	プラグ開口 335 には、インクパック 7 の後端に設けられたプラグ 72 が設けられている。(2, 5)

出力では“faces”が正しく訳出されていないため誤訳であるが、一方の評価者は評価 5 としている。

入力	第 1 熱媒体経路 232 はその内部に熱媒体としての水を通流させる。
正解	Water as the heat medium flows through the inside of the first heat medium passage 232 .
出力	The first thermal medium path 232 flows through the water as a heat medium.(1, 4)

この例では入力と正解とで主語が変化しており、それゆえ入力が使役文であるのに対し、正解は能動態の文になっている。一方の出力は主語は入力と一致しているが、能動態で訳してしまっているため翻訳としては誤りとなっている。

入力	同軸芯線 52a は、導線で構成される。
正解	The coaxial core 52 a is constituted by a conducting wire.
出力	The coaxial core 52 a is constituted by a conductor. (2, 5)

conductor だけでも導線という意味があるため、出力も正しいと考えられるが、評価者の一方が評価 2 としている。

3.2 「重要情報」の定義の違いによる評価の差

特許庁の基準では、「重要情報」がどれだけ正確に伝達されているかが評価ポイントになるが、「重要情報」の定義は評価者の主観に委ねられている。以下に示す例ではこの定義の評価者間のズレが影響している可能性がある。

入力	For example, FIG. 1A shows a plan view of a videoconference room with a typical arrangement.
正解	例えば、図 1A は、典型的な構成からなるビデオ会議室の平面図である。
出力 1	例えば、図 1A は、典型的な構成を有する会議室の平面図を示す。(3, 5)
出力 2	例えば、図 1(a)は、一般的な配置の会議室の平面図を示す。(3, 5)
出力 3	例えば、図 1a は、典型的な構成を有する会議室の平面図を示す。(3, 5)

この例ではすべてのシステムにおいて「ビデオ」という部分が抜けているが、一方の評価者は評価 5 としている。これは「ビデオ」の部分がさほど重要ではないと判断し、ほかの重要情報が全て含まれているため評価 5 としている可能性もある。

入力	In this example, assume that jobs have been deleted and A′=12 is obtained.
正解	今回の例ではジョブが削除され A′=12 となったとする。
出力 1	この例では、ジョブが削除され、A′=12 であると仮定する。(3, 5)
出力 2	この例では、ジョブが削除され、a=12 が得られると仮定する。(3, 5)

この例では情報の不足はないように見えるが、一方の評価者は「仮定する」と訳していることが誤りと判断している可能性がある。

入力	The recording medium 212 is a memory card freely removable from, for example, the camera body 200.
正解	記録媒体 212 は、例えばカメラ本体 200 に着脱自在になされたメモリカードである。
出力	記録媒体 212 は、例えばカメラ本体 200 から着脱自在に着脱自在のメモリカードである。(3, 5)

この例では出力に重複が見られるが、評価者の一方が評価 5 としている。確かに重複があっても重要情報が正しく伝達されていると考えられなくもない。特許庁基準では過剰に出力された情報をどう扱うかの基準が設けられていないため、このような評価割れが起こる可能性がある。

入力	第 1 熱媒体経路 232 はその内部に熱媒体としての水を通流させる。
正解	Water as the heat medium flows through the inside of the first heat medium passage 232.
出力	The first thermal medium path 232 flows through the water as a heat medium.（1, 4）

この例では入力と正解とで主語が変化しており、それゆえ入力が使役文であるのに対し、正解は能動態の文になっている。一方の出力は主語は入力と一致しているが、能動態で訳してしまっているため、文全体の意味としては誤りとなっている。ここで、文全体の意味が誤っているから評価 1 とするか、個々の名詞は正しく訳出されているから評価 4 とするかで差が生まれていると考えられる。次に示す例も同様である。

入力	また、その固体撮像装置を用いた電子機器を提供することを目的とする。
正解	Also, it is desirable to provide an electronic system using the solid-state imaging device.
出力	It is also an object of providing an electronic device using the solid state imaging device.（1, 4）

この例でも述部の訳が少しズレているが、必要な部品は全て正しく訳出されている。このような差は、例えば否定の有無といった場合にも起こりえる。

3.3　訳語の不統一

専門用語などが入力文中で複数回出現した際にその訳が統一されていない場合や、より適切な訳がある場合などに、評価が割れる傾向があるようだ。

入力	For information, new items are not limited to added items but include changed items.
正解	なお、新規項目は、追加された項目に限らず、変更された項目を含む。
出力 1	情報のために、新しいアイテムは、追加項目に限定されず、変更項目を含む。（3, 5）
出力 2	情報については、新しいアイテムは、追加項目に限定されず、変更されたアイテムを含む。（3, 5）

“item” という単語の訳が「アイテム」であったり「項目」であったりと訳が揺れている。

入力	In a particular embodiment, the photo sensor is a photo-multiplier tube.
正解	具体的な一実施形態ではそのフォトセンサは光電子増倍管である。
出力 1	特定の実施形態では、光センサは光マルチプライヤー管である。（3, 5）
出力 2	特定の実施形態では、フォトセンサーは photo-multiplier チューブである。（1, 4）

“photo” が「フォト」や「光」と訳されていたり、photo-multiplier tube が様々に訳されている。決まった訳語がなさそうな場合には、一般的には専門用語全体で統一して和名もしくは洋名（カタカナ）で訳すべきであり、和名と洋名が混ざって訳されることは好ましくないと考えられる。

入力	FIG. 15 is a representation of one unit of a carboxymethyl cellulose molecule.
正解	図 15 は、カルボキシメチルセルロース分子の 1 つの単位の模式図である。
出力	carboxymethyl セルロース分子の一単位の表示である。（4, 2）

この例では英単語がそのまま日本語文に訳出されている。頭字語など英単語をそのまま訳出すれば良い場合ももちろん存在するが、日本語訳が存在する場合には日本語にするべきであると考えられる。しかしながらこの例のように、英単語のまま訳出されても意味はわかると考えることもでき、基準を決めないと評価が割れる要因になる。

入力	糖タンパク質 D は、いくつかの宿主受容体のうちの 1 つを認識し、結合し得る。
正解	Glycoprotein D recognizes and can bind to one of several host receptors.
出力	Sugar proteins D may recognize and couple one of several host receptors.（3, 5）

この例では「糖タンパク質」が “Sugar proteins” と訳出されているが、これは誤訳で “Glycoprotein” が正しい訳である。この例からは、専門用語の正しい訳を知っていないと正しく評価できないということと、現在の翻訳システムが専門用語を構成的に翻訳してしまうことの弊害が見て取れる。

3.4　一文だけでは訳が確定しない

稀なケースではあるが、分脈がないと訳が確定できな

いものもあった。

入力	The encryption method in this case is AES as mentioned above.
正解	この際の暗号化方式は、上述のように AES である。
出力 1	この場合の暗号化方法は、上述のような AES である。(3, 5)
出力 2	この場合の暗号化方法は、上述したような AES である。(3, 5)

この例では、AES というものが 1 つしかないならば（前の文脈でそのような記述がされているならば）正解訳とする必要があるが、AES には何種類かあり、そのうちの一つが前の文脈で記載されているならば、出力が正しい訳となる。このように、文脈情報がないと正しく翻訳できないものも、割合は少ないが存在する。

4 まとめ

本稿では WAT2017 の特許日英・英日翻訳タスクの JPO Adequacy Evaluation に対して、2 つの観点からの分析を行った。質の良い翻訳システムの評価において κ 係数が不当に低く算出されることを明らかにし、κ 係数だけを見ても評価の信頼性は測れないことを指摘した。今後は κ 係数だけでなく、評価値の詳細も明らかにする必要があると思われる。

また評価結果が 2 名の評価者間で乖離している例を分析したところ、その要因は評価者の単純なミスによるもの、評価尺度の定義の解釈の違いによるもの、訳語の不統一によるものなどがあった。評価者のミスはある程度は仕方がないが、評価尺度の定義や訳語の不統一の扱いなどは事前に基準を決めておくなどすれば、より一致度の高い評価が行えると思われる。

参考文献

[1] Overview of the 4th Workshop on Asian Translation, In Proceedings of the 4th Workshop on Asian Translation (WAT2017).
[2] https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm