機械翻訳における誤りの傾向

渡辺桂子　建石由佳

工学院大学情報学部
Email: j106123@ns.kogakuin.ac.jp,yucca@cc.kogakuin.ac.jp

1. はじめに

現在、機械翻訳システムは、ウェブで公開されたものや、ソフトウェアパッケージとして販売されたものがみられるが、翻訳された文の意味が通じないことが多く、その精度は決して高いとは言えない。そこで、改善点を見つける目的で、機械翻訳の誤りを調べることにした。

ここでは、翻訳精度に影響すると考えられる機械翻訳の誤りには訳語選択の誤りと構文の誤りがあると仮定し、品詞ごとに 2 種類の誤りの傾向を調べることにした。

「日英新聞記事対応付けデータ（JENAAD）」[1]を使用し、コーパスでの訳文と原文を 2 種類の機械翻訳システム[2, 3]で翻訳した結果を市販の英和辞書[4, 5]の訳と対照することにより正解かどうかの判定を行った。

2. 原文と訳文の単語単位での対応表の作成

訳語選択の誤りを調べるためには、原文と訳文の単語レベルでの対応が必要であり、構文の誤りを調べる際にも原文の構文を訳文でどのように表しているかを調べるため、各英単語の訳語との対応が必要である。しかし、JENAAD での原文と訳文の対応付けは文単位であったため、単語単位での対応表を人手により作成した（これを「単語単位での対応表」と呼ぶ)。ただし、英単語と日本語の単語が 1 対 1 対応するとは限らず、1 つの英単語に対し、複数の日本語の単語が対応することもある。

手順は、以下の通りである。

① Tsuruoka Tagger[6]を使用し、JENAAD の英文を単語分割し、品詞付与を行う。

② JENAAD の英文にコーパス上で対応している日本語文(以下、この訳文を「正解文」という)と機械翻訳ソフト 2 種を用いた翻訳結果に対して MeCab[7]を用いて形態素解析と品詞付与をする。

③ 上記①,②の英文と日本語文の単語を人手により意味で対応させる。助詞などの付属語は対応する英単語がない場合がある(主格の「が」など)が、その場合は直前の英単語と同じ単語に対応させる。

対応させた結果を Microsoft Excel の表形式データとした。

3. 正解判断基準の設定

訳語選択と構文について以下のような基準を設定した。

3.1. 訳語選択

単語単位での対応表を用い、英単語ごとに対応する機械翻訳結果の単語と正解文の単語を比較し、機械翻訳結果の訳語選択が正しいかどうかを調べる。

訳語が正しいかどうかの判断は言葉の微妙なニュアンスにより難しいため、今回は、一般に使われる辞書を用いて判断した。しかし、必ずしも辞書が正しいとは限らないため、2 種類の辞書を使用している。

訳語選択が正解かどうかの判断は以下のような手順に従って行う。

① 正解文の単語と機械翻訳による訳語を比較し、正解文と機械翻訳結果が同じであれば機械翻訳の訳語は正しいと判断する。

② 正解文と機械翻訳結果が異なる場合は、辞書

を引く。どちらかの辞書で正解文の訳語と同じ意味上の分類項目に機械翻訳結果の訳語が載っていれば正解とし、正解文の訳語が1つあるいは両方の辞書に載っているが機械翻訳結果の訳語がない場合、あるいは、異なった項目に載っている場合は誤りとする。

③ 正解文の訳語がどちらの辞書にも載っていない場合には、正解判断が不可能とする。

3.2. 構文

いくつかの品詞の語は、文の構造に対して重要な役割を持つ。たとえば、英語の前置詞、日本語の助詞などである。このような語に対しては、原文での係り先を翻訳文でも反映させなければ翻訳後の意味が変わってしまうと考え、原文での係り先と翻訳文での係り先を比較した。

調査手順は、以下の通りである。英文での単語と正解文と翻訳文での単語の対応は、「単語単位での対応表」を使用する。

① IN、CC、TO、RB、RBR、RBS[1]に対して、人手作業によりそれぞれの英単語の係り先を調べる。また、IN、CC、TOについては項(argument）の範囲も調べる。

② 訳語選択で作成した英文の単語と翻訳文での単語の対応を基に、①で行った係り先、項の範囲に対応する日本語の単語を調べる。ここでは、日本語文での単語の文法的役割は考慮しない。

③ 正解文と機械翻訳結果の文を日本語係り受け解析器 CaboCha[8]で解析する。CaboCha は、文節で判断するため、単語単位の対応表と対応が異なるが、文節の中に該当する単語が含まれていたら正解とする。

④ ②での日本語文での係り先、項の範囲がCaboCha の解析結果と係り受けが同じかを比較する。Cabocha の解析結果が曖昧な場合は両方正解とする。

4. 機械翻訳の誤り率

JENAAD より 25 文に対して調査した。1 文の平均は 12.4 単語である。

4.1. 訳語選択

調査対象品詞は VB, VBD, VBG, VBN, VBP, VBZ, NN, NNS, NNP, NNPS, RB, RBR, RBS, JJ, JJR, JJS, PRP, PRP$, DT の全19品詞とした。対象は 313 単語であった(表 1 参照)。

表 1 では、判断不可能とした英単語も、判断不可能となった理由を調べるため、誤りとして数を数えている。JENAAD 行の「誤り」とは正解文の単語が辞書に載っていなかった数を意味する。

特徴として、名詞は、誤り率としては他と比較すると低いが、各文での名詞の割合を平均するとおよそ 3 割であるため、誤りの総数としては多くなっている。代名詞は、JENAAD では、ほとんど訳されていなかったのに対し、機械翻訳では訳されていたため、判断不可能となる英単語が多く、訳語選択の誤り率が高くなっている。冠詞は、基本的には日本語に翻訳されていないが、誤りの 1 単語は、原文で単独で現れた”This”を DT と判定しているものを、正解文では「この(連体詞)こと(名詞)」、機械翻訳ではどちらも「これ(名詞)」と翻訳したものである。

4.2. 構文

調査対象品詞 IN, RB, RBR, RB, CC, TO の全 6 品詞、311 単語に対して調査した。正解文と機械翻訳結果の文との比較は行わず、英文での係り受けが日本語文でも同じように表現されているかを調べた(表 2 参照)。

JENAAD 行は、正解文が意訳されたために英文との違いが生じることや、CaboCha の解析誤りが考えられるため、このような結果となっている。

[1] これらは Penn Treebank[9]における品詞を示す。

表 1 訳語選択の品詞別誤り数と割合

	VB,VBD, VBG,VBN, VBP,VBZ (動詞)		NN,NNS, NNP,NNPS (名詞)		RB,RBR, RBS (副詞)		JJ,JJR,JJS (形容詞)		PRP,PRP$ (代名詞)		DT (冠詞)	
全単語数	49		147		6		48		16		47	
JENAAD	15	0.31	19	0.13	2	0.33	7	0.15	11	0.69	1	0.02
コリャ英和	19 (6)	0.39	35 (19)	0.24	3 (1)	0.50	9 (2)	0.19	11 (0)	0.69	1 (0)	0.02
ピカイチ	24 (10)	0.49	32 (14)	0.22	4 (2)	0.67	12 (6)	0.25	11 (0)	0.69	1 (0)	0.02

※各カラムの左側は対象品詞の誤り語数、右側は（対象品詞の誤り語数/対象品詞の総語数）
※誤りに判断不可能とした英単語を含む
※カッコ内は、判断不可能とした英単語を除いた単語数

表 2 構文の品詞別誤り数と割合

	IN (前置詞)		RB,RBR,RBS (副詞)		CC (接続詞)		TO	
全単語数	61		4		17		8	
JENAAD	8	0.13	0	0.00	5	0.29	2	0.25
コリャ	16	0.26	1	0.25	7	0.41	2	0.25
ピカイチ	12	0.20	1	0.25	2	0.12	0	0.00

※各カラムの左側は対象品詞の誤り語数、右側は（対象品詞の誤り語数/対象品詞の総語数）

4.3. まとめ

以上の結果を踏まえ、誤りの割合を”誤り単語数/調査対象数”として計算した(表 3 参照)。

正解文の訳語が辞書に載っていない訳語を使用している割合は、17%であった。機械翻訳で、確実に誤りと判断された訳語選択の割合は、どちらのシステムも 10%前後であり、調査で使用した 1 文の平均は 12.4 単語のため、およそ 1 文につき 1 単語の誤りが発生しているということである。

表 3 機械翻訳における誤り率

	訳語選択	原文の構文解析
JENAAD	0.17	0.17
コリャ	0.25(0.09)	0.29
ピカイチ	0.27(0.10)	0.17

※カッコ内は、判断不可能とした英単語を除いたとき

5. 品詞別の傾向

品詞別に以下のような傾向があると考えられる。

動詞は、判断不可能とされた単語の割合が多い。これは、翻訳する際には、助動詞、be 動詞などと組み合わせて翻訳されるため、特に be 動詞などは、対応するものがないことや、翻訳の仕方によって訳語が異なることがある。しかし、今回は、正解文や辞書を基準に判断したため、誤り率が高くなってしまったことが考えられる。

名詞は感覚的には訳の誤りが目立つが、実際には誤りの割合は少ない。しかし、文の中に占める割合が多く、また、物事の名称であるため、誤ると大きく意味が異なってしまい、訳語選択の少しの誤りでも目立つのではないかと考えられる。

機械翻訳が副詞の構文を誤ったのは、vigorously が前の単語を修飾するケースであった。他の副詞では、後ろの単語を修飾しているのに比べ、このケースのみ前にある単語を修飾していた。副詞は、後ろを修飾すると決めてしまっている可能性があり、副詞の係り受けは、ルールを増やす必要があるかもしれない。ただし、ここでは数が少ないため、もう少し調べていく必要がある。

接続詞は、誤りが多いが、今回は曖昧なケース

を両方正解とした[2]ので実際はさらに誤りが多い可能性がある。一方、翻訳文での並列の判断がCaboCha の解析エラーによるものもあった。さらに、CaboCha による接続詞の解析エラーが他の品詞の係り受けにも影響することがありうるため、実際よりも誤り率が高く出ている品詞もあるとも考えられる。

表 2 にはないが、名詞句自体の範囲と係り先は正しいが、名詞句の中での係り先と項の範囲が違っているものがあった。例えば、”an early conclusion of a peace treaty”は、” of a peace treaty”が” an early conclusion”と判断できるが、機械翻訳では、” of a peace treaty”が”conclusion”に係り、”an early”は、” conclusion of a peace treaty”に係るように翻訳されていた。

6. おわりに

調査を始める前は訳語選択のほうが難しく、誤りが多いと考えていた。しかし、今回の結果では、訳語選択と構文との誤り率はあまり大きくは変わらなかった。また、構文の誤りの判断では曖昧なケースが判断に影響した。しかし、これは英文の構造だけでは判断が難しく、意味を見ることによって正解がわかるものであり、このようなものの判断は、機械翻訳では難しいのではないかと感じた。

今回の調査では、すでにあるソフトウェアを使用しながら行ったため、人手で判断の揺れがなく、さらに時間も短縮される。しかし、使用したソフトウェアの精度が誤り率に影響する可能性が考えられたため、その内容を分析する必要がある。

参考文献

[1] Masao Utiyama and Hitoshi Isahara. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72-79.

[2] 「コリャ英和!一発翻訳 2010 for Win」ロゴヴィスタ株式会社

[3] 「翻訳ピカイチ 2009 アカデミック版 for Windows」株式会社クロスランゲージ JAN コード 4947398097662

[4] 「研究社　新英和(第 7 版)中辞典・和英(第 5 版)中辞典 」

[5] Yahoo!辞書「プログレッシブ英和中辞典第 4 版」http://dic.yahoo.co.jp/

[6] Yoshimasa Tsuruoka and Jun'ichi Tsujii, Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, Proceedings of HLT/EMNLP 2005, pp. 467-474. http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/postagger-1.0.zip

[7] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of EMNLP-2004, pp.230-237. http://mecab.sourceforge.net/

[8] cabocha-0.53.exe http://chasen.org/~taku/software/cabocha/win/

[9] Marcus M.P., Santorini B., Marcinkiewicz M.A. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, Vol.19, No.2, pp. 313-330.

[2]例えば、”world peace and security”は(world (peace and security))とも、((world peace) and (security))とも解釈できるが、”世界の平和と安全”（曖昧になる）でも、”世界平和と安全”でもどちらでも正解とした。