機械翻訳システムの評価と改善

田中康仁
兵庫大学
E-mail:yasuhito@humans-kc.hyogo-dai.ac.jp

〔０〕はじめに

機械翻訳システムが日本に於いて本格的に研究開発されはじめて約２０年弱の年月が経過した。しかし、機械翻訳には色々な問題点がある。ここでは機械翻訳の現状を概観し品質向上にあたっての問題点、方法を検討する。

〔１〕機械翻訳システムの現状

機械翻訳システムの現状を分析するには次のような項目を調べなければならない。

１）機械翻訳システムの開発会社と製品の種類
２）研究開発を続けている企業
３）専門用語辞書の種類と量
４）価格
５）翻訳速度
６）市場占有率
７）機械翻訳システムの版（ヴァージョン）

特に次にあげる項目は重要である。

８）品質

機械翻訳システムで最大の課題は品質である。まだまだ十分なものには到ってはいない。これは開発者の問題であると同時に利用者からの問題提供、翻訳の質に対する系統的なクレームの申立てによる改良にあると思われる。利用者が商品に対して申立てを行うのは当然のことであるがこれが充分行われていないのではないだろうか？雑誌等の評論記事はソフトウェア会社や開発企業の宣伝に振りまわされていて本当の姿が表現されていない。

日本電子工業会を中心にした企業の研究会で、機械翻訳の機能テストを行っているが、プログラムテストの観点からは有効であるが、機械翻訳システムが持たなければならない知識の量については充分な検査とはなっていない。

今では機械翻訳システムにどの程度有用な知識データが入っており、それが活用されているかということを調べる段階にきている。

数文程度の比較ではなく数万件程度のテストデータを常に準備する。しかも、半年毎とか、１年毎にそれらを更新して、新しいテスト文で検査して、改善を計らねばならない。日英、英日の検査を行うことを考えると、パラレル・コーパスの作成が必要である。

消費者センターと協力し、品質の向上に努めることが重要である。

機械翻訳システムには限界が明確に述べられていない。また、どれだけの文でどのようにテストしたかについても述べられていない。これは利用者の誤解と失望をまねく。

さらに必要なことは、今後どのような方針で改定をするかということも書かれていない。閉じた完全なシステムであるならばよいが、自然言語のようにオープンな世界で利用するものには記述が必要である。

次に現在売られている機械翻訳システムを大量のデータを用いて評価してみる。

〔２〕どのようにしてテストデータを作るか？

テストデータを作るにあたっては次の三つの方法が考えられる。

１）大量のコーパスを分析する。
２）既に出版されているＣＤ－ＲＯＭの利用
３）多勢の人によるテスト文の作成

等が考えられる。次にこの個々の方法を詳細に述べる。

1)の方法として英語の新聞その他のデータ・ベースが機械可読媒体として売られている。これらは研究等については特別の許諾書に著名をすれば利用可能である。このようなものは多くの場合ＣＤ－ＲＯＭで売られている。量としても適当である。

２）の方法として既に出版されているＣＤ－ＲＯＭを利用する方法としては

英和辞典、和英辞典として売られているＣＤ－ＲＯＭをＤＤＷＩＮ３２等の検索ソフトを利用し、例文を抽出し利用することも考えられる。英文のビジネス文作成のためのＣＤ－ＲＯＭが売られている。これらを利用するのも一つの方法である。

しかし、著作権の問題も考えなければならない。

３）の方法としてあげた多勢の人々によるテスト文の作成は我々のように大学で教えているものにとってはデータ収集のよい方法である。

１００人の学生のタッチ・タイプの練習として英文、日本語文の対を入力させる。約３００文を入力する。前期、後期では２回の機会があり、３万文×２＝６万文（英⇔日の文）が機械可読媒体で集まる。しかし、学生の雑多な文をうまく整理する方法を考えなければならない。あまり複雑な入力方法では誤った文ばかりになる。

〔３〕評価の方法について

評価については色々な方法がある。幾つかのものを次に示す。

(1)機械翻訳システムを対象に、Ａ，Ｂ，Ｃのランクを付けるための評価
(2)機械翻訳システムの利用と使わない方法での費用の分析による評価
(3)品質向上のために各機械翻訳システムにはどのような問題点があるか、大量のデータによるテストにより問題点を明確にする。どのような改善をすれば品質が向上するかを評価する。
(4)その他

等の方法がある。

ここでは(3)の方法を基にして考えてみる。しかし、この方法は問題点の指摘であるため、開発者にとってはあまり良い気持ちを持てない。

しかし、問題点を明確にすることが、品質向上に結びつくことを理解していただきたい。

〔４〕機械翻訳システム用評価データ

機械翻訳システムを評価するにあたって重要な点はどのような評価データを集めるかということである。ここでは日本電子化辞書株式会社（ＥＤＲ)の英文コーパスを用いることにした。この英文コーパスはタグ付けされているが、このタグを取り除いた文を用いた。これを単語数別に分類する。単語数別文の数は次の通りである。

単語数	1	2	3	4	5	6	7	8	9
文の数	0	19	460	1,889	5,030	6,798	7,791	8,416	8,698

10	11	12	13	14	15	16	17	18	19
9,114	9,018	9,176	9,050	8,815	8,446	8,245	7,466	6,559	5,283

20	21	22	23	24	25	26	27	28	29
3,645	562	464	321	235	142	77	46	23	12

30	31	32	33	合計
2	0	0	1	125,803

このコーパスの中で？（疑問符）と！（感嘆符）の付いている文の数を単語数別に調べると次のようになる。

単語数	2	3	4	5	6	7	8	9	10	11
？の文数	0	50	314	872	1,105	1,092	965	811	566	475
！の文数	1	20	29	45	34	34	17	23	19	16

12	13	14	15	16	17	18	19	20	21	22
322	251	206	156	101	72	71	42	38	5	3
10	21	7	5	6	2	1	4	1	0	0

23	24	25	26	27	28	29	30	33	合計
0	0	2	1	0	0	0	0	0	7,652
1	0	0	0	0	0	0	0	0	296

疑問文は６単語のものが最も多く、全体的に他の文と比べ短い。感嘆文も５単語のものが最多である。

ＥＤＲの英文コーパスを分析した結果、次のような特徴がある。

文の総数	125,670文	
最長の文	33単語の文　1文	
最短の文	2単語の文 19文	
平均単語数	12.22単語	
単語数の中央値	12単語	
疑問文の数（？の数）	7,652文	6.09%
感嘆文の数（！の数）	296文	0.23%

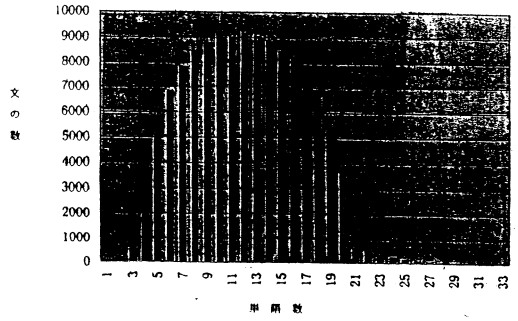
また単語数が２０以上になると急に少なくなっている。さらにEDRコーパスには次の特徴がある。

(1) 新聞の文に比べ人称代名詞が主語になる文が多い。

(2) 疑問文、感嘆文の数が多い。

(3) 単語数が２１単語以内のもので９９％を占めている。

(4) 単語数が２０以上がきわめて少なくなっているのは、ＥＤＲコーパス作成時の何らか操作を行ったと思われる。



〔５〕テスト手順

次のような手順で機械翻訳システムをテストする。

１）７単語までの文で約２万件のデータが得られる。このデータを機械翻訳システムにかけて翻訳する。

２）翻訳結果に対して良否の評価を行う。１～５点の点数を付ける。

３）点数別、単語数別に分類する。同一点数で翻訳に悪い影響を与えている原因コードを付ける。複数の原因のコードがある場合は１つだけ原因コードを付ける。

例えば、専門用語を追加すれば良くなるとか語と語の共起関係を追加すれば修正可能であるとかいう原因コードを付け修正にまわす。

４）原因コード点数別に分類する。

このようにし翻訳システムの改良を系統的に行うことができる。さらに、問題点の原因コード別に集計し、統計をとり、品質向上の参考資料を得ることができる。

〔６〕評価の具体的方法について

(1) ソフトウェアとしての安定性

機械翻訳システムはコンピュータ上で動くソフトウェアである。このためどのようなデータに対しても、ソフトウェアが何の応答もなく終了するということは避けなければならない。

Ａ社のシステムでは、ある特定のデータを処理させると、ソフトウェアが何の応答もなく終了していた。しかし、Ａ社は直ちに新しい版（ヴァージョン）を出してしまい、今では問題なく処理するシステムになっている。

(2) 問題点だけの収集

Ｂ社のシステムは構文解析、その他の処理で処理不能となると、一定の記号列を表示し、訳語の単語列を表示する処理を行い、次の文の処理に移ってしまう。このため大量のデータを処理してみると、多量の誤りが出現した。そこでＥＤＲの英文コーパスの７単語以内の約２万文の処理中に１，２２６文見つかった。これらを開発者に送ったところ、次のような返信を得た。

1）ＥＤＲの英文コーパスの中に英文字の誤り、編集上の誤りがあった。これは我々データ作成上の問題点であった。

2）省略形

例えば、you're のようなものについての対応が不十分であることが判明した。

また、文の一部が省略されているものもある。これについては対応されていない。

Yes, I certainly will (do)　do の省略

3）一部の文法上の欠点があることが、開発者によって認められた。倒置文に対して弱いということである。

例　Behind every caltural image exists those taboos.

以上のことは開発者が誤った処理によるデータを検討すれば、ただちに判明するものであった。これは誤った処理を約１，２００文集めたからである。約５％の文が処理できていない。誤って処理したデータを分析する中で、一文毎の個

別の問題か、それらの中にある共通の問題であるかが判明した。

(3) 訳文の理解度による評価

機械翻訳システムでは「理解容易度」による評価と「忠実度」による評価がある。ここでは理解容易度により評価方法を行ってみた。

評価に用いた英文はＥＤＲの英文コーパスを採用し、単語数が２～７のものを用いた。採点は英語の教員が行った。

評価結果　Ｃ社の英日翻訳システム

	5	4	3	2	1	合計	平均値
2 単語文	15	1	1	1	1	19	4.47
3 単語文	336	97	16	11	0	460	4.65
4 単語文	1,369	336	157	25	2	1,889	4.61
5 単語文	3,655	808	510	53	4	5,030	4.60
6 単語文	4,742	1,379	595	81	1	6,798	4.58
7 単語文	4,471	2,331	870	118	1	7,791	4.43
合計	14,588	4,952	2,149	289	9	21,987	4.53
%	66.35	22.52	9.77	1.31	0.04	100%	

評点の高いものが良い結果を示す。

平均値をみると４．５程度で大変良いように思える。しかし、評点５のものだけを○とし、他のもの　は×とすると１００点満点で平均66点となる。これは我々の感覚とほぼ同じであることがわかる。

次にＤ社、Ｅ社の製品についても同様の方法で評価してみる。

評価結果　Ｄ社の英日翻訳システム

	5	4	3	2	1	合計	平均値
2 単語文	6	9	4	0	0	19	4.10
3 単語文	259	127	54	14	6	460	4.34
4 単語文	1,099	551	167	67	5	1,889	4.41
5 単語文	2,555	1,680	645	150	0	5,030	4.32
6 単語文	3,585	2,087	953	174	0	6,798	4.33
7 単語文	2,222	3,541	1,638	368	22	7,791	3.97
合計	9,725	7,995	3,461	773	33	21,987	4.21
%	44.23	36.36	15.74	3.52	0.15	100%	

評点の高いものが良い結果を示す。

評価結果　Ｅ社の英日翻訳システム

	5	4	3	2	1	合計	平均値
2 単語文	8	5	4	2	0	19	4.00
3 単語文	232	151	58	19	0	460	4.30
4 単語文	966	634	228	60	1	1,889	4.33
5 単語文	1,645	2,104	1,047	228	6	5,030	4.02
6 単語文	1,907	2,965	1,598	328	0	6,798	3.95
7 単語文	2,149	3,575	1,667	400	0	7,791	3.96
合計	6,907	9,434	4,602	1,037	7	21,987	4.01
%	31.41	42.91	20.93	4.72	0.03	100%	

評点の高いものが良い結果を示す。

このＣ社はＥＤＲプロジェクトで中心的役割をはたした会社である。単語数が少ない文であるためか、非常に良い翻訳結果が得られた。しかし、単語数が多くなると少しずつ平均値が下がっている。

Ｄ社の製品はＣ社に比べ劣っていることがわかる。Ｅ社の製品はまだまだ改良しなければならないことがわかる。

この方法では、翻訳結果の良い文と悪い結果の文を分けることがなされるので、悪い結果を集め分析することにより、改良方法がわかる。個々の文固有の問題点、知識データの不足、文法体系、文解析等の問題であるかを判別することができる。

言語学者や辞書学者、語用論学者へ機械翻訳で処理できないデータや問題点を提供し、研究が発展することを期待する。

また単語数別の分析は、問題点の単純化をはかり、原因をつかみやすくする。長い文に発生する固有の問題点はみつけだせない。翻訳ソフトウェアのような大きなシステムは、一度に全てを調べあげることはできない。個々の単純な部分を分析し、誤りを修正し、より複雑な誤りをみつけだすようにすべきである。

Ｃ社以外の数社について試したところ、Ｃ社より平均点で約１点程度低いことがわかった。この１点の差は大きな差である。今後の改良を期待する。

Ａ社、Ｂ社、Ｃ社とも、この論文を作成した時点から版（ヴァージョン）が更新され少しずつ良くなっている。このような実験を行うと、このデータについては改良がなされるため、さらに新しいテストデータを準備しなければならない。新しい多量の言語データを標準化しなければならない。どのような条件があれば標準化されたデータかは大きな課題である。最近では大量のWWWが作られているのでこの英文を集め、単語数別に整理し、同一の文はまとめ、頻度をつけてテストデータとして利用することもできる。

〔７〕この評価方法について

単語数別ファイルを作り機械翻訳システムを評価する方法は、問題点の抽出が容易で興味ある方法である。

しかし、翻訳結果の評点付けは人間の作業であり、大変労力のかかる作業であった。この作業を自動的に行うように機械翻訳システムに組込んでおくことが重要である。最終的な判断は人間の作業であるが、ある程度のところまでは機械的に可能である。これにより翻訳結果の大まかなグループ分けが可能である。このようにして作業の迅速化が図られる。

例えば

１）構文解析で曖昧さが減らすことができない。

２）未知語が出現した。

３）意味解析のパターンが無い。

４）専門用語が無いため合成訳を作成した。

５）並列処理の解析がうまく行えなかった。

等機械翻訳システムで誤る場所は幾つかある。これらの問題点について重み付けを行い、評点を付けるのも一つの方法である。自動的評価システムを作らなければならない。

〔８〕機械翻訳システムの改良についての提案

前述までのように機械翻訳システムをコーパスを用いて、単語数別に分類することにより短い文から順次評価する方法を提案した。この経験より、次のような改良方法を述べる。

１）大量のコーパスを用いて、機械翻訳により結果の良くないデータを見つけ出す。

結果の良くないデータを20,000文程度集めるとすれば、３０％の翻訳ミスが有る機械翻訳システムでは約６万７千文のコーパスを準備すればよい。

どのような分野の内容を重点にしたコーパスを準備

すればよいか考えなければならない。

2）利用者からのクレームとデータの収集

利用者から翻訳できない文を集めテスト・データとする。又は、機械翻訳システムの特別ユーザ（モニター）からのデータ収集

3）単語についてのデータ収集

コーパスや辞書の見出し語や、ＷＷＷから英単語や日本語の単語を抽出し、機械翻訳辞書の見出し語と照合し不足のデータを補う。

4）複合語の補強

我々、日常生活の中で使われている複合語を集め、機械翻訳用辞書と照合し、不足しているものを補う。複合語は日本語と英語のように対訳になったものでなければならない。また対訳語が複数ある場合は、使用条件も付けなければならない。集めた複合語がどの程度機械翻訳システムの辞書と一致したか、一致しなかったものはどの程度かを測定しなければならない。また追加される複合語がどの程度実際に使われているかも調べねばならない。

筆者は日本語の複合語として四文字漢字列、五文字漢字列を数十万語整理して持っている。これらの中で頻度が高いものに訳語を付けるのも一つの解決方法である。

5）専門用語について

専門用語については専門用語辞書を作成している出版社や、ソフトウェアー会社や、学会や協会と協力し、購入することが得策と考えられる。

機械翻訳システムに組み込まれているもの以外を購入すべきである。このためには購入時に内容を評価してくれる第三者が必要である。

6）慣用表現の補強

慣用表現の補強も重要なテーマである。我々は約3万件の慣用表現を集めた。日本語と英語が対になったものである。

7）例文の収集

このごろの機械翻訳システムは用例ベースの機械翻訳機能が付け加えられている。このための例文を集めることが重要である。

我々は学生のタッチタイプの練習として日本語、英語の対を入力させている。これを定期的に集め、重複を整理し、誤りを修正すれば良い例文データとなることが判った。一年間に約6万文程度の日本語と英語の対を集めた。これを単語数別に整理し、1～12単語程度の文が用例ベースの機械翻訳機能に有効であることが判った。

8）結合価文法の文型データの収集

結合価文法の文型パターンは慣用表現の収集中に、動詞句の中からかなり抽出できることが判った。また、1）で述べた、大量の翻訳誤り文の中から特定の動詞に注目し、検索し、ＫＷＩＣを作成することにより資料が得られる。

9）その他

その他、色々な文法上の改良については1）の大量のコーパスを翻訳した結果の中から、分析するのが良い方法である。

1）～9）まで改良の方法についての指針を示すことができた。これらを実行し、新しい版を作り、何％の翻訳結果の改善が得られるか、得られたかの評価をしなければならない。

〔９〕おわりに

機械翻訳システム（英⇒日）のテストを行い評価を行った。大量の英文データを単語数別に分析し、整理し、単語数の少ないものから順次テストするという方法を考えついた。これにより機械翻訳システム（英⇒日）の品質の向上をはかる一つの方法ができた。

さらに色々な方法を考えてみたい。これら試行から、大量の日英対訳付コーパスを作成しなければならないこともわかった。また、市販されているＣＤ－ＲＯＭを機械翻訳のテストに利用するのも一つの方法である。

実際に稼働している機械翻訳システム上で2万件強のテストデータで翻訳を行ってみた。数社の協力が得られた。

さらに、評価ばかりでなく改良のための指針が具体的に得られた。

〔１０〕参考文献

(1) 安田賀計　らくらく使えるビジネス文書1,230文例　ＣＤ－ＲＯＭ　日本経済新聞社

(2) 田久保浩平、橋本光憲　英文ビジネスレター文例大辞典　ＣＤ－ＲＯＭ　日本経済新聞社　15,000文

(3) 塩澤　正、スコット シェフェルバイン　インターネット英語表現辞典　三修社　1998年1月

(4) 新編英和活用大辞典（自然な英文を書くための３８万例）ＣＤ－ＲＯＭ版　研究社　1996年

(5) 社 日本電子工業振興協会「自然言語処理システムの動向に関する調査報告書」 平成９年４月

(6) 牧野武則　評価技術「機械翻訳」Ｂｉｔ別冊　共立出版　1988年９月

(7) 長尾　真　「機械翻訳はどこまで可能か」岩波出版　1986年６月

(8) Language and Machines:Computers in Translation and Linguistics, National Academy of Sciences National Research Council (1966)

(9) 高橋文子　３語で通じる英会話　明日香出版社　1998年５月

〔１１〕データについて

英文データは日本電子化辞書㈱の英文コーパスを利用した。

日本電子化辞書㈱のプロジェクトに参加した企業がＥＤＲ英文コーパスをあまり利用していない。翻訳システムにもっと活用してほしいものである。

ＥＤＲコーパスには日本語のコーパスもある。これを日英機械翻訳システムに利用すべきである。

この論文の翻訳評価基準について

理解容易性の詳細な内容については参考文献(7) Ｐ５５を参照されたい。

この原文は Language and Machines: Computer in Translation and Linguistics, National Academy of Sciences-National Research Council (1966) .（ALPACレポート）の中にある。