冗長な補助文による伝達度のシミュレーション

長　瀬　久　明
（兵庫教育大学）

統計的機械翻訳は多くの言語を翻訳の対象にできる。しかし、その翻訳精度は未だ十分ではない。本研究は伝達度を改善するために補助文を考え、簡単なシミュレーションを行った。その結果、補助文は翻訳精度が余り良くない言語ペアにおいて、とくに有効に働くことが示唆された。

キーワード：機械翻訳，補助文，伝達度，シミュレーション

長瀬久明：兵庫教育大学大学院・行動開発系教育コース、総合学習系教育コース・教授，〒673-1494 兵庫県加東市下久米942-1，
E-mail: hisac@hyogo-u.ac.jp

A Simulation of Transmission by Using Redundant Supplementary Sentences

Hisaaki Nagase
（*Hyogo University of Teacher Education*）

The statistical machine translation can aim at a lot of languages for translation. However, translation accuracy is not yet sufficient. In this study we thought about supplementary sentences to improve the transmission level and made a simple simulation. As a result, the followings were suggested. In the language pair with good translation accuracy, the supplementary sentences improve the transmission rate. In the language pair at intermediate translation accuracy, the supplementary sentences improve the transmission rate most effectively. In the language pair at poor translation accuracy, the attainment to the level that reads easily is difficult though the supplementary sentences improve the translation rate.

Key Words: Machine Translation, Helper Statement, Communication Level, Simulation

Hisaaki Nagase: Professor, Hyogo University of Teacher Education, 942-1 Shimokume, Kato-city, Hyogo 673-1494 Japan.
E-mail: hisac@hyogo-u.ac.jp

1章　はじめに

今日、誰でも無料でインターネット上の機械翻訳システムを利用することができる。先に提供されたシステムは文法的機械翻訳という方式であった（たとえば[1]）。続いて、統計的機械翻訳方式によるシステムも提供された（たとえば[2]）。統計的機械翻訳は開発費用が資金的にも時間的にも文法的機械翻訳に較べて激減する[3]。このメリットを生かして統計的機械翻訳は積極的に、多くの言語を翻訳対象としてきた。google 翻訳の対象言語の数は2011年現在、約60に達している。しかし、どちらの方式も誤訳を無視できる精度には達していない。統計的機械翻訳の誤訳について黒田らは、文献３（p.35）で次のように述べている。

「統計的機械翻訳の出力は翻訳がうまく行った場合と行かなかった場合の落差が非常に大きい。うまく行かなかった場合、しかも、それは決して稀ではないのだが、統計的機械翻訳の出力は有意味な表現どころか、容認可能な表現にすらならない。」

機械翻訳の研究においては翻訳精度が評価尺度であるが、林田らは翻訳精度に代わる評価尺度としてインタラクティビィティ（相互作用性）を提案している[4]。また、山本もこれに似た尺度として伝達度と、伝達度を高める具体的な一方法として「補助文」を提案し、日韓翻訳を例に実験している[5]。伝達度とは補助文を含めて伝達できる割合をいう。

本研究の目的は山本のアイデアに基づいて、補助文を様々な翻訳精度の言語ペアに適用した場合をコンピュータ・シミュレーションし、伝達度への寄与について考察することである。

2章　伝達度のシミュレーション

2.1　原文とその補助文の例

次のような３つの日本語文を考えてみよう。

「今日は天気がいい。」

「野田氏が決選投票で海江田氏を逆転した。」

「ハリケーン「アイリーン」は勢力を弱めたが、27人が死亡したほか、約500万戸が停電した。」

次に、文の意味を保ち、表現を変えた文を考えよう。表現を変えるとは、格の変更、同義語(類似語)への変更、語順の変更、態の変更、短文化、などである。このような文を山本は補助文と言っている[5]。次の文は上の日本語文の補助文の例である。

「今日はいい空模様だ。」（格の変更、類似語への変更）

「決選投票で野田氏が海江田氏を逆転した。」（語順の変更）

「決選投票で海江田氏は野田氏に逆転された。」（態の変更）

「ハリケーン「アイリーン」は力を弱めた。しかし27人が死んだ。ほかに約500万戸が停電した。」（短文化）

「ハリケーン「アイリーン」の勢力は弱まった。しかし死亡者が27人出た。ほかに停電が約500万戸あった。」（格の変更、類似語への変更、など）

2.2　コンピュータ・シミュレーション

言語 a と言語 A のペアは機械翻訳できるとしよう。言語 a の文が n 文、a_i（i＝1,…,n）あり、どの文にも補助文が２つ、すなわち、b_i、c_i（i＝1,…,n）が作れたとしよう。実際は、補助文を作りやすい文、作りにくい文があるが、コンピュータ・シミュレーションは元もと、複雑な現象を極限まで単純化して隠れた傾向を明らかにするために行うものである。今は最初のシミュレーションであり、補助文の有無と伝達度の関係を明らかにすることに主眼を置いている。補助文の作り易さについては、今は考えないことにする。

次に、文 a_i が「aA 翻訳」された訳文 A_i は、元の意味を保持している場合（正しく翻訳された場合）か、あるいは、元の意味を失った場合（誤訳された場合）か、どちらか一方であるとしよう。ここでも勿論、実際の翻訳結果はほとんどの場合、この両極端の中間にある。しかし、このコンピュータ・シミュレーションは「補助文の有無と誤訳が伝達度に及ぼす影響」だけを考えている。

そこで、正しく訳される割合を x（０＜x＜１）、誤訳される割合を y（０＜x＜１）とすると x＋y＝１である。また補助文 b_i、c_i についても同様とする。つまり、ある言語ペアでは、どの文も一定の割合 y で誤訳されるとする。実際は、原文 a_i が誤訳されやすい構文や語彙を含んでいる場合、補助文 b_i, c_i も誤訳されやすい傾向がありうる。この傾向を無視することは危険であるが、もし考慮すれば簡単な計算は不可能になる。そこで、「誤訳が独立して起きるような補助文の作り方」という課題を設定しておき、現在のところは、誤訳は文ごとに、お互いに独立に発生するものと仮定する。

また逆に、実際は、読者が誤訳を修正し正しく解釈できるような、都合の良い場合もある。前後の文はお互いに関連があるからである。しかし、今は単純に、各文は単独に解釈され、訳文 Ai が誤訳されると伝達できないと単純化する。

ただし、正しく訳されたか誤訳されたかの判別は、正しくできると仮定しておく。誤訳された場合、多くの場合意味が取れず、たとえ意味が取れても、前後の文との意味の関連が著しく低くなる、というのが一応の理由づけである。しかし、原文と異なる意味をもつ訳文が出力され、そちらの意味が正しいと判断を誤る危険はあり得る。その危険が補助文により減ることは期待できるが、皆無になりうると即断は出来ない。もし皆無になれば、完璧な異言語コミュニケーションが実現したことになる。

2.3 具体例

数値例で伝達度をシミュレーションしてみよう。一つの段落の長さの程度として、14の文を考えよう。

原文、a_1,…, a_{14} がある。各文は y＝0.5で誤訳されるとする。このレベルは、正しく翻訳された文があれば、次の文は誤訳といった、読みづらいレベルであろう。補助文が無い場合、14の訳文中、約7文が伝達できない。

次に、各文に二つの補助文がある場合を考える。

a_1,…, a_{14}. b_1,…, b_{14}. c_1,…,c_{14}.

読み手は、おそらく、

a_1, b_1, c_1, …, a_{14}, b_{14}, c_{14}.

の順に読むことになる。ただし、誤訳は文ごとに、お互いに独立に発生するものと仮定した。誤訳される割合 y が0.5なら、正しく翻訳される割合 x も0.5である。ここで、i 番目の文 a_i と、その補助文 b_i, c_i とが、すべて誤訳である割合は、$y^3=0.5^3=0.125$ となる。逆に、3 文中少なくとも 1 文が正しく翻訳される割合は 1 −0.125＝0.875である。誤訳かどうかは正しく判断できると仮定したので、各文は 8 割以上、伝達できる。全14文中では、伝達できない文の数は、14×0.125 ＝ 1.75 となり、14文中約 2 文弱が伝達できない。補助文が無い場合、14文中約 7 文が伝達できなかったことを考えれば、大きく改善されている。

2.4 誤訳率と伝達度の関係

誤訳率 y が0.5で、補助文の数が 2 つの場合について前節で計算した。誤訳率 y と補助文の数がその他の場合についても、前節と同様の式で計算できる。そこで、誤訳率が0.05～0.95、補助文の数が 0 ～ 7 の場合について伝達度を計算し、誤訳率を横軸に、伝達度を縦軸にとり、補助文が 0 ～ 7 の場合についてグラフを描く（図 1 ）。図 1 から、言語ペアの誤訳率により、補助文の効果が異なることが分かる。すなわち、

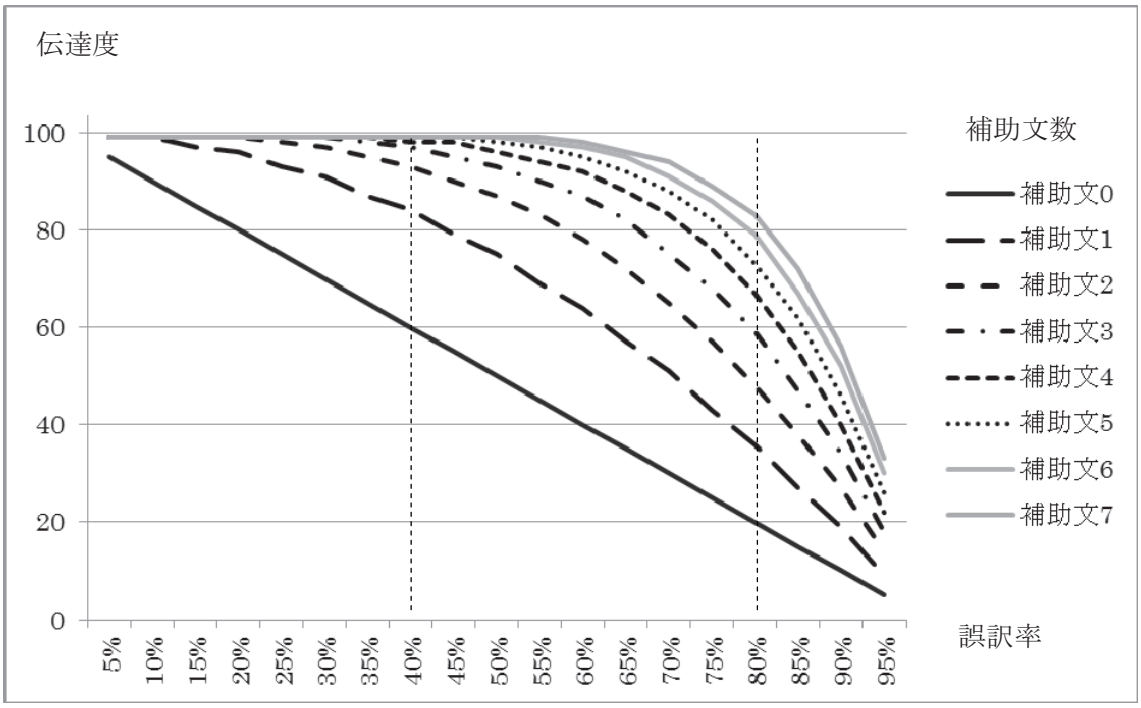

図 1　誤訳率、伝達度、および補助文の数の関係

① 誤訳率が 5 ～40％の、翻訳精度が比較的良い言語ペアでは、1 ～ 2 つの補助文は伝達度を改善し、90％以上の十分な伝達度が得られ、それ以上の多数の補助文は、伝達度を改善する効果は小さい。

② 誤訳率が40～80％の、翻訳精度が中間的な言語ペアでは、1 ～ 2 つの補助文は伝達度を改善する効果が最も大きい。また、3 ～ 4 つの補助文があれば、伝達度をさらに改善する効果がある。

③ 誤訳率が80％以上の、翻訳精度が比較的低い言語ペアでは、6 ～ 7 という多数の補助文があっても、伝達度80％以上を達成することは容易でない。

3章　考察

日韓翻訳は文法的翻訳の精度が元もと、90％以上なので①の範囲である。山本は、被験者が、確認効果がある補助文と、必要性を感じない補助文があると評価したと

報告している[5]。この結果は①と一致している。

補助文はむしろ②のような、伝達度が中間レベルの言語ペアで、より大きな寄与が期待される。機械翻訳の精度がこのレベルの言語ペアは多いと思われる。

発信者は補助文を作ることが求められる（図２左）が、補助文の作成とは母語の言い換えであり、発信者は母語以外の言語習得を求められない。

さらに、補助文を簡単なルールで作ることが出来れば、ソフトによる作成も可能になると期待される（図２右）。

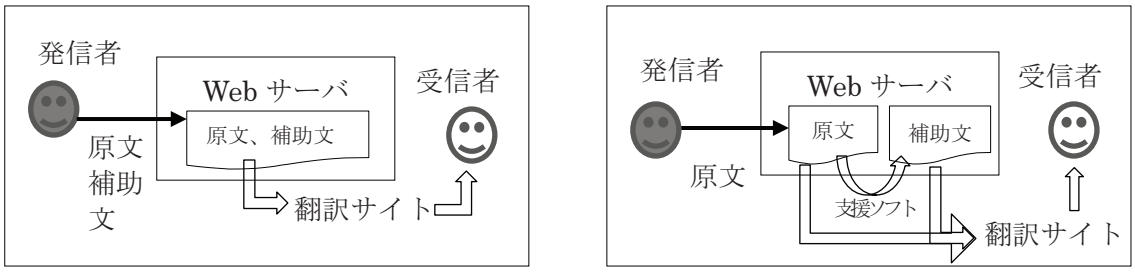

図２　補助文のＷｅｂへの応用（左：手作業、右：支援ソフト）

４章　あとがき

発信者は補助文（書き換え文）を作成する必要がある。可能な書き換え方は翻訳元の言語によって異なる。例えば「語順の変更」は日本語では可能であるが、不可能な言語も多い。また、翻訳先の言語に関する知識があれば、ある書き換え方が有効かどうか、確認できる。いったん、十分な種類の書き換え方が見いだされると、これに従うことで、翻訳先言語に関する知識がない人も、有効な補助文を作りやすくなる。さらに、多くの翻訳先言語に対して有効であれば、それに越したことはない。さらに、単純な書き換え方ほど、ソフトウェアによって自動的に作成できる可能性が高まる。この段階に至ると、書き換えは非常に楽になる。

また、受信者には、翻訳の正誤を判断し、誤訳を捨てるという負担がある。この点では、言語ペアの翻訳精度が高まるほど、少ない数の書き換え文で読めると思われ、読みにくい文に出会ったとき、追加の書き換え文を表示できるような仕組みが求められるだろう。

また、本研究では考えなかったが、異言語コミュニケーションには次のような問題点がある。

Ⅰ　ある言語（文化、あるいは民族）では、ある意味のことを言うとき、ある言い方をする。しかし、この言い方を別の言語に翻訳したとき、翻訳先の文化、あるいは民族には、その言い方が無い、あるいは、しない場合がある。

Ⅱ　表現（明示）せず、暗示する習慣になっていて、状況あるいは文脈に伝達を任せる場合がある。どのような内容を暗示するかは、言語の特徴による。

これらは、同言語コミュニケーションや、通訳、習得の場合は考える必要がないが、機械翻訳では顕在化する。コミュニケーション（意思疎通）における本質的な問題点といえる。

参考サイト、参考文献

[1] http://honyaku.yahoo.co.jp/

[2] http://translate.google.co.jp/

[3] 黒田　航, 加藤鉱三：今の機械翻訳に利用者が望めること，望めないこと，日本語学，p.40，vol.28-12，2009.

[4] 林田尚子, 石田　亨：翻訳エージェントによる自己主導型リペア支援の性能予測, 電子情報通信学会論文誌 D-I Vol. J88-D-I No.9 pp.1459-1466, 2005.

[5] 山本歩：機械翻訳を利用したグローバルな交流を助ける冗長な補助文に関する研究, 兵庫教育大学2007年度修士論文.

（2011．８．31受稿，2011.11.28受理）