英日機械翻訳システムＴＯＭＣＡＴＳ メンテナンスのためのデザイン

原田雅弘
日本アイ・ビー・エム株式会社ナショナル・ランゲージ・サポート

1.はじめに

機械翻訳プログラムが対象とする自然言語は非常に多様であり、デザインを開始した時点でそのすべてのパターンを想定しておくことは不可能である。特に、利用者の観点から見た細かな要求は、実際にシステムを運用し、利用者からの声を集めない限り、検討を行う事はできない。

そこで、これらのさまざまな文や要求に対応していく必要があるが、対象となる文や要求は膨大な数となる。これに対応するには、次の事が要求される。

1. 迅速なエラーの発見
2. 迅速な分析
3. 迅速な修正

本論文は、これらの点について、英日機械翻訳システム TOMCATS がどのようにデザインされているかを示したものである。

2.英日機械翻訳システム TOMCATS

機械翻訳システム TOMCATS は日本アイ・ビー・エム株式会社の NLS (National Language Support) で研究、開発が行われている。NLS は、主に各国から送られて来た英文のマニュアルなどを日本語に翻訳する部門であり、TOMCATS の実用化に際し、最も大きな利用者となる部門でもある。

TOMCATS の研究、開発がこの部門で行われているのは、より多くの翻訳サンプルを得るとともに、実際の利用者たる翻訳者からのシステムに対する要求をより多く、より直接的に得、システムの改善を図っていくためである。

ここでは、これらの要求に迅速に対処し、機械翻訳システムの品質を上げていくために、TOMCATS がどのように設計されているかを示していく。

3.エラーの分類

翻訳エラーは次のように分類される。

1. 訳が無い
2. 訳が日本語になっていない
3. 原文の品詞認定などに誤りがある:li, 原文の構造分析（並列構造、係り受け等）に誤りがある
4. 訳語選択に誤りがある
5. 日本語の文体が不適当である
6. その他

「文体」は、利用者の好みによって左右されることも多く、対応が困難であることも多い。しかし、このような利用者の要求に対応していかない限り、機械翻訳システムは実用的なものとはならない。

4.エラーの収集

エラーの収集には大きく分けて次の２つがある。

1. 翻訳者の要求
2. 研究開発担当者による分析

文体の問題などは利用者の要求から得る事が多い。しかし、利用者の要求を集めるのは、利用者の余分な負担を求める事でもあり、実際に大量の問題点を収集するには、研究開発担当者による分析が中心となる。

研究開発担当者による分析のために、問題となる文章（または問題かも知れない文章）をシステム

TOMCATS - An English to Japanee Machine Translation System, - Design for Maintenance
Masahiro Harada
1623-14, Shimo-tsuruma, Yamato-shi, Kanagawa, 242 Japan
National Language Support, IBM Japan Ltd.

が自動的に検出するプログラムを提供している。このプログラムは次のような事例を検出する。

1. 翻訳処理の失敗
   翻訳処理コンポーネントのいずれかが処理に失敗し、異常な訳を出したか、またはまったく訳を出さなかった。
2. 利用頻度の低い文体
   例えば、技術文書では倒置法が用いられる事はない。
3. 利用頻度の低い品詞
   例えば、計算機に関する技術文書では、"can" はたいてい助動詞として用いられ、「缶」の意味で用いられる事はない。
4. 曖昧な語の用法
   例えば、条件節接続詞として用いられる "as" や "while" 、助動詞の "may" などはその意味するところが曖昧な場合が多い。ただし、文体などから、一意に用法を特定できる場合は除く。

上記のような場合、プログラムはそれぞれに応じたメッセージを提供する事ができる。

実際には、上記のプログラムによって検出されないエラーの方が多く、翻訳結果を担当者がチェックする必要がある。この作業を助けるために、上記とは逆に、プログラムによって「チェック不要」のメッセージを付けることもできる。次の場合について、チェック不要のメッセージが返され、プログラムによってチェックの対象から外される。

1. 全大文字語、タグのみ等、処理が不要なもの
2. 全体が１つの名詞または複合名詞として辞書に登録されている
3. 上記の大文字語、名詞＋
   "command"/"display" 等

最近のテストでは、「翻訳の失敗」と「使用頻度の低い文体／品詞」を合わせて約 5%、「チェック不要」が約 20% となった。

その他の分析支援ツールとして以下のプログラムが提供されている。

1. 翻訳結果比較プログラム
   同じファイルに対して行った２つの翻訳結果を比較する。適当な期間をおいて翻訳された２つの翻訳結果を比較して訳が変更された文をリストし、その変更を検討する。
2. 例文検索プログラム
   特定の単語またはその組合せを持つ例文を検索する。英日の対訳で出力も可能（ただし、例文は少数）

5.エラーの原因の分類

翻訳エラーの原因は次のように分類される。

1. 原文の文法ミス
2. プログラム／ルールの誤り
3. 辞書項目の不足:li, 辞書の誤り
4. 機械的には処理困難な文
5. 機械的には処理困難な文体の要求
6. その他

6.エラーの分析／修正のための人員

翻訳システムが運用の段階に有る場合、エラーの発見分析、修正の全てを研究開発担当者が行う事はできない。そこで、これらは次のように役割分担される。

1. 利用者
   エラーの発見
2. 翻訳担当者
   エラーの集計、辞書の問題の同定
3. 辞書担当者
   辞書の修正
4. 研究開発担当者
   上記の担当者から報告される問題の詳細分析および修正

上述のように、研究開発担当者自身が行う事もある。この場合は、エラーの収集から分析、修正までをすべて行う事になるが、辞書の修正については、辞書担当者と相談の上で行う事になる。

7.TOMCATS のデザイン

TOMCATS のデザインにあたっては、翻訳システムのメンテナンスを考慮して次のいくつかの基本方針を定めた。

1. 強力な辞書
   訳語の選択以外にも、さまざま曖昧さの解消を辞書情報を使って行い、担当者が訳の選択だけではなく、翻訳の構造的な選択までも含めた変更を容易に行えるようにした。
2. 独立したコンポーネント構成
   各コンポーネントの処理を独立したものとし、失敗の原因の発見、修正を容易にする。
3. マクロ言語の設定
   プログラムしやすいマクロ言語を設定することにより、修正を容易にする。

4. 各種のモニターの充実
翻訳エラーの分析ツールの作成

8.TOMCATS のコンポーネント

1. 解析前処理
複合名詞の処理、書替え規則による自動書替えを行ない、解析の高速化、曖昧さの事前解消を行う。
2. 英文解析
英文の構造解析。名詞以外の複合語処理
3. 解析木選択
複数の解析木から１つを選択する。
4. 係り受け選択
並列構造の決定、前置詞や不定詞等の係里受けなどの決定を行う。
5. PAS (Predicate Argument Structure) 変換
解析木を PAS に変換する
6. 英文 PAS 変形
英文 PAS を日本語に変換しやすい形に変形する。
7. 英日トランスファー
英文 PAS を日本文 PAS に変換する。
8. 日本語 PAS 変形
日本文 PAS を自然な日本語を生成できるよう変形する。
9. 日本語生成
日本文 PAS から日本語文字列を生成する

9.モニター・プログラム

次のようなモニター・プログラムが提供されている。

- 各コンポーネントの入出力表示
各コンポーネントについて、入力および出力を表示する。
- コンポーネント別モニター
各コンポーネントごとに次のようなモニター・プログラムが提供されている。
- 解析前処理
置き換えられた単語列とそれに与えられた日本語訳のリストを示す。
- 英文解析
適用がテストされたすべてのルール番号、または適用されたすべてのルール番号を表示する。
インタラクティブ・モードでは、解析木のノード属性の表示、部分単語列からなる部分木の表示、さらにこの部分木に特定のルールが適用できるか、適用失敗の場合はルールのどの部分で失敗したかを示すこともできる。
- 解析木選択
すべての解析木を表示し、それぞれの解析木について、選択ルールから返されるメッセージとそのスコア、および合計スコアを表示する。、
- 並列構造／係り受け選択
解析木中の並列構造／係り受けの候補を表示し、それぞれについて、適用される選択ルールとルールから返されるメッセージとスコアを表示する。係り受けの選択を行うべきノードが複数ある場合には、１つの選択が行なわれるごとに選択後の木を表示する。
- 英文 PAS 変形
PASTRAN モニター（後述）
- 英日変換
単語に複数の訳語が与えられている場合に、辞書のデータと各エントリーとそれぞれに与えられたスコアを示す。一部、PASTRAN で書かれた部分については PASTRAN モニターが利用できる。
- 日本文 PAS 変形
PASTRAN モニター
- 日本語生成モニター
述部の生成の際に利用される情報を示す。一部、PASTRAN で書かれた部分については PASTRAN モニターが利用できる。

PASTRAN モニターは次のモニター出力を出すことができる。

- 適用された全てのルールのリストと適用された PAS 構造のトップ・ノード名
- 適用された全てのルールの入力および出力
- 特定のルールについての入出力。ルールの適用に失敗した場合は、その失敗したルールの部分

この他に、PLNLP や LISP/VM で標準として提供されるモニター・プログラムが利用される。

10.エラーの分析手順

TOMCATS には、「エラー分析手順書」が提供されている。これに従ってエラーの分析を行っていくと次のことが割るようになっている。

1. エラー分析の手順
2. モニター・プログラムのどれを使うべきか
3. モニター・プログラムの使い方

4. モニター・プログラムの出力の注目点
5. 修正が可能であるか
6. どの辞書／データ／ルール／プログラムを修正するのか
7. 修正の為のプログラムの使い方
8. 修正の結果の確認の方法
9. 副次効果の検討方法
10. 修正の権限
11. 一時的修正と永続的修正の選択指針

辞書やデータの一時的な修正委関しては、すべての担当者が修正可能になっている。しかし、永続的に影響を及ぼし修正については、その修正権限はその責を持つ担当者に限られている。

辞書などの修正に関しては、翻訳対象のファイルやマニュアルだけを対象とする一時的修正と、分野別辞書経野修正、システム辞書の修正などのレベルに分けられる。

11.コンポーネントと辞書データ

TOMCATS の各コンポーネントがどのような辞書情報を参照しているかを下に示す。これらの辞書はほとんどユーザーに解放されており、いくらかの教育を行えば、ユーザー自信で修正を行う事ができる。また、これらの辞書のほとんどはユーザー単位（正確には各ユーザーのディレクトリ単位）でほじる事が出来る。

12.おわりに

ここに示されたように、TOMCATS はユーザーに解放された辞書のみの修正で大きくシステムの改善を行えるように設計されている。また、プログラム内のモニター・プログラムや、支援プログラムによって、問題点の発見、分析、修正が素早く行えるようになっている。

これらの仕組みを用いて、TOMCATS は翻訳業務や研究開発担当者のチェックを通して大量の改善要求などを収集し、これに対応し、システムの改善が行なわれている。

ffl 参考文献 “

1. Harada M,日本アイ・ビー・エムにおける機械翻訳システムの設計と運用(1992, MT Fair 93)

	Pre-process	English Analysis	Tree selection	Reatt-achment	Transfor-mation	E-J Transfer	Japanese Gen.
Analysis rules		O					
Analysis dictionary							
PRIORITY		O	O				
JEM data				O			
Others		O	O	O		O	
Transfer dictionary							
Verb			O	O	O	O	O
Adjective			O	O	O	O	O
Adverb			O			O	O
Preposition			O	O		O	O
Noun	O			O		O	O
&a. rules					O	O	O
Other tables	O				O	O	O