[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

Automatic Evaluation of Output Quality for Machine Translation Systems

Yu Shiwen
Institute of Computational Linguistics
Peking University

Abstract

Automatic evaluation of output quality for machine translation systems is a difficult task. The Institute of Computational Linguistics of Peking University has developed an automatic evaluation system called MTE. This paper introduces the basic principles of MTE, its implementation techniques and the practice experiences.

Introduction

In China Machine Translation ought to make great contributions to social development. Therefore machine translation was one of the national key projects in the past five years (1986-1990). Automatic evaluation of output quality for machine translation systems is a subtask of the machine translation project.

Automatic evaluation of translation quality for MT can realize standardization and rapidity of fixed quantity testing. It is of great significance for promoting research in machine translation and its applications. The Institute of Computational Linguistics of Peking University has developed an Automatic Evaluation System for Machine Translation called MTE. In this direction, we made the first firm step.

1 Examination Model

In general, there are two types of examination models. One is a subjective test model, another is an objective test model. The subjective test model is shown in fig.l.

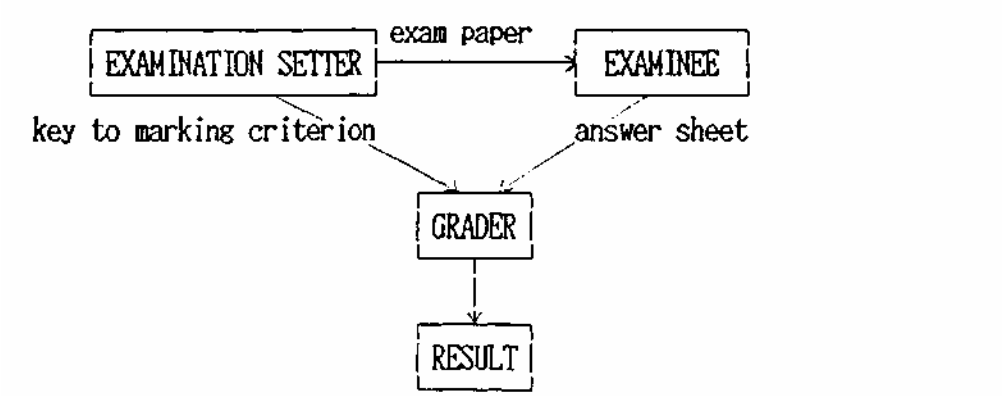


fig.1. Subjective Test Model

The objective test model(multiple choise) is shown in fig.2.

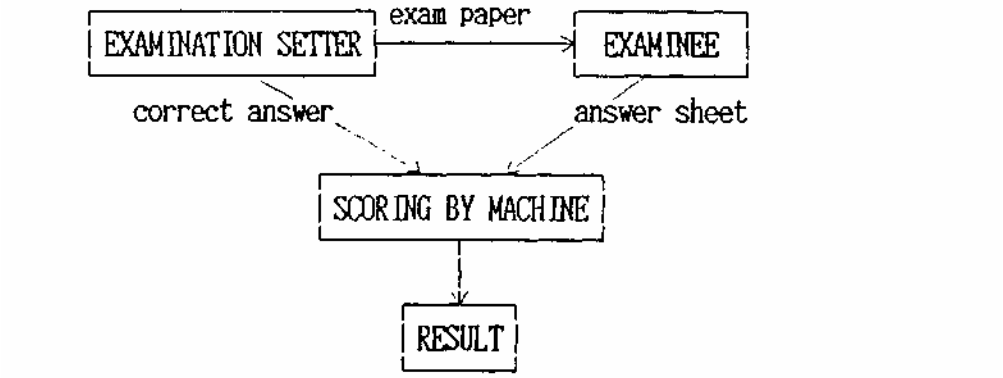
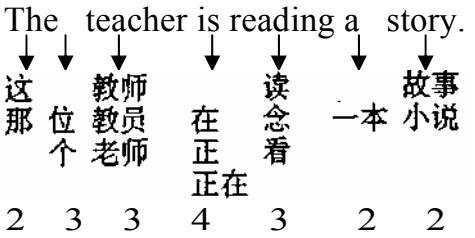


fig.2. Objective Test Model

The examinee makes a choice of one from 4 given answers when he takes part in the objective test. Automatic scoring is easily realized by computer for the objective test. But automatic evaluation of output quality for MT systems is more difficult than the test of monolingual knowledge. The corresponding relation between an English sentence "The teacher is reading a story" and its Chinese translations is shown below.



The numerals in the last line above represent the numbers of possible Chinese words corresponding to every English word in the sentence. In general, classifiers, such as "位" or "个", should be added between "这" and "教师" in Chinese. That is, there are at least 2x3x3x4x3x2x2=864 acceptable translations for this simple English sentence. The objective test model is useful to shape our idea and method of automatic evaluation for MT output quality. The automatic evaluation model is shown in fig.3.

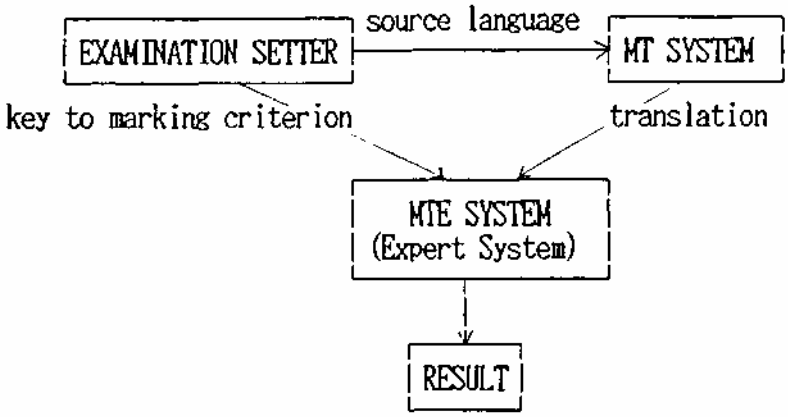


fig.3. Automatic Evaluation Model for MT

An MT system can be regarded as an expert system. The MTE system can also be regarded as an expert system.

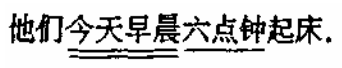
2 Basic Principles of MTE

1. We will take English-Chinese machine translation as a concrete example to introduce MTE, though its principles and implementation techniques are independent of specific language-pairs.

2. According to the existing level of MT, the evaluation of translation quality is made on single sentences.

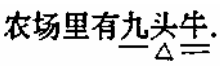
3. The test points are extracted from Chinese translations of English sentences. Some examples of test points are given below.

Ex.1 They got up at six this morning.



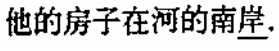
test point: sequence of "time words".

Ex.2 There are nine cows on the farm.

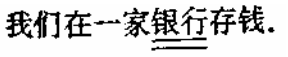


test point: usage of the classifier "头".

Ex.3 His house is on the south bank of the river.



We keep our money in a bank.



test point: multiple meanings of the word "bank".

4. To cover the language phenomena in machine translation as widely as possible the test points should be distributed reasonably over all categories.

For example,

vocabulary
idioms
polysemantic words
irregular verbs
usage of classifier in Chinese
comparative degree
superlative degree
adjusting word order
processing of undefined words

basic patterns of simple sentences
usage of"不"or"没"in negative sentences
tense
complex sentences (object clause,...)
emphasis
ellipsis
mood
.
.
.
etc.

5. Aiming at the test points, a vast amount of normal source language (SL) sentences had to be chosen and made into an SL file. The Target Language (TL) file is formed by translating the sentences in the SL file.

6. The translation quality test program in the MTE system not only gives all the test points' scores, but also makes comprehensive evaluation with statistical methods for the translation files.

At present, there are 3,200 carefully selected English sentences in the SL file. There are two sets of translation files in MTE. There are several hundred test points in the MTE. These test points are divided into 6 classes: words, idioms, morphology, elementary grammar, moderate grammar, advanced grammar. A weight is assigned to every class for comprehensive evaluation.

class	number of original sentences	weight
words	315	0.1
idioms	317	0.1
morphology	772	0.2
elementary grammar	724	0.2
moderate grammar	874	0.3
advanced grammar	191	0.1

Table 1. Number of sentences and weight of the classes.

3 TDL: Key of Implementation Techniques

The key of MTE is how to describe different translations and the marking criterion of the specified test point for an English sentence. To describe these demands a description language for testing, called TDL, has been designed. TDL is a context-free formal language. We explain the meaning and the usage of TDL with concrete examples.

Ex.1 They got up at six this morning.

```
score    R→(492:1)*$A[的]$B$C*
  subject  R→(492:0)*
           $A→早晨/上午
complement $B→六/6
           $C→点钟/点/时
           ##
```

test unit

This test unit contains 2 test items and 3 complementary declarations of non-terminals $A, $B, $C. "R→" is a start sign of every test item. Subject 492 is a code of a test point. "*" can match an arbitrary character string in the translation. The content in the brackets "[" and "]" is optional. If an MT system gives an output such as Ex.1 of section 2, then subject 492 can win one point.

Ex.2 There are nine cows on the farm.

```
R→(364:1)*&(nine)头&(cow)*
R→(364:0)*
```

The acceptable Chinese translations of "nine" and "cow" can be found in English-Chinese dictionary of MTE. The representation with "&" can reduce the definitions of non-terminals. The representation can also be used in random substitution of the specified word in an English sentence with those words of the same type.

Ex.3 His house is on the south bank of the river.
We keep our money in a bank.

```
R→@x
##
R→*银行*
    (R→(309:2)@x:*岸*)
R→*
##
```

There is an evaluation rule for testing multiple meanings of a word in MTE. In Ex.3 the two sentences must be tested jointly. If the two corresponding Chinese words of “bank” in the two English sentences are entirely correct, then the subject 309 wins two points, else subject 309 gets nothing. The sign @x is a variable and @x stores the former translations. It's content is referred in later unit.

4 Syntax formulas of TDL

The main syntax formulas of TDL are written in extended BNF form.

: : = , | , {,} ,<,>

are symbols of the meta-language, not symbols of TDL.

The formula

<A::={<B>}

is equal to

<A>::=<null>|<B><A>,

where "|" represents "or".

<test file>::=<test unit>##{<test unit>##}

<test unit>::=<test item>|<complement><test item>|<test item><complement>|
<complement><test item><complement>

<test item>::=R-><scoring><match pattern>|
R-><scoring><match pattern><test item>|
R-><scoring><match pattern>(<test item>)

<complement>::=<non-terminal>-><type>{/<type>}{<complement>}

<scoring>::=<null>|(<subject>:<mark>{,<subject:mark>})

<match pattern>::=<external match>|<internal match>

<external match>::=<match element>{<match element>}

<internal match>::=<implicit element:<external match>
{<implicit element>:<external match>}

<match element>::=<arbitrary element>|<definite element>|<optional element>|
<implicit element>

<arbitrary element>::=*

<definite element>::=<type>

<optional element>::=[<type>]

<implicit element>::=@<numeral>|@<small letters>

<type>::=<terminal>|<non-terminal>|<reference>

<terminal>::=<Chinese word>|&[<English word cat.](English word)|
#<attribute of Chinese word>

<non-ternimal>::=$<capital letters>

<reference>::=@<implicit element>

<attribute of Chinese word>::=<range of length><Chinese word cat.>[<subcat.>]
{,<semanteme>}

<range of length>::=<null>|<min length>:<max length>

<null>::=

5 Structure of MTE

The structure of MTE is shown in fig.4.

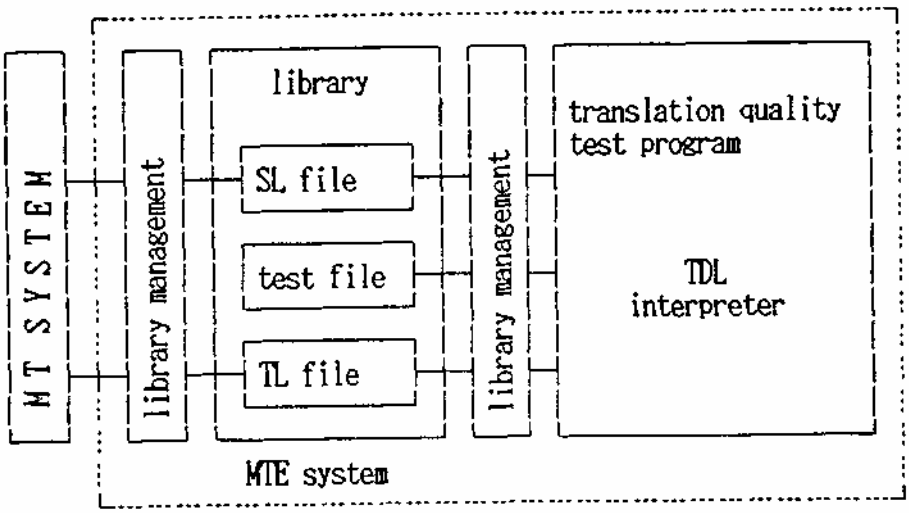


fig.4. Structure of MTE

The library is composed of SL file, TL file and test file. At present the test file contains 12,000 records. The TDL interpreter is the kernel of the translation quality test program in MTE. The translation quality test program is coded in PASCAL. The library management program is coded in DBASEIII.

6 Practice of Automatic Evaluation

MTE has tested the qualities of two sets of translation files. One set it translated by a person whose English level is not very high. Another is translated by a Machine Translation system called "TRANSTAR". The "TRANSTAR" system is developed by the China National Software & Technology Service Co. (CS&S). The designer of the "TRANSTAR" system is professor Dong Zhangdong.

The results of testing are listed in table 2.

class	full scores	weight	translation by "TRANSTAR"			translation by human		
			actual scores	%	weighted scores	actual scores	%	weighted scores
words	315	0.1	301	95.6	9.6	256	81.3	8.1
idioms	317	0.1	279	88.0	8.8	271	85.5	8.6
morphology	772	0.2	584	75.7	15.1	692	89.6	17.9
elementary grammar	733	0.2	525	71.7	14.3	638	87.0	17.4
moderate grammar	1791	0.3	1143	63.8	19.2	1429	79.8	23.9
advanced grammar	424	0.1	178	42.0	4.2	308	72.6	7.3
comprehensive evaluation					71.2			83.2

Table 2. Comparison between Translations by Machine and Human

Note: The above results are based on the data received in August, 1990.

The results tally with the actual situation of the translations.
Time required to test each set of translation files is about 13 minutes on an IBM PC/AT.

7 Goals for Further Improvement

1. To increase the quantity of the library
   We plan to increase to 10,000 English sentences.
2. To improve the quality of the library
3. To enlarge the fields of evaluation

   - title
   - text

4. To deepen the depth of evaluation
   We want to evaluate the output quality of machine translation based on context analysis.

5. To implement standardization

6. To extend to other language pairs

   - Japanese-Chinese
   - Chinese-English
   - German-Chinese

8 Acknowledgement

My colleagues Jiang Xin, Zhu Xuefeng and professor Hou Fang of Heilongjiang University made important contributions to the development of MTE.
I have benefited from advices of professor Ma Xiwen of our institute.
I would like to thank professor Dong Zhendong of CS&S and other friends.
They supported and helped our research work.

References

[1] *Language and Machines,* "Computers in Translation and Linguistics". National Academy of Sciences, National Research Council (1966).

[2] B. Henisz-Dostert, R.R. MacDonald & M. Zarechnak, *Machine Translation,* Mouton, (1979).

[3]長尾真，国際翻訳技術フォーラム報告，電子工業月報，第３１巻 第６号，PP 35-41 (1989).

[4]長尾真，機械翻訳文の質の評価と言語の制限，情報処理 Vol.26, No. 10, PP 1197-1202, (1985).

[5] 牧野武則，評価技術，ｂｉｔ，１９８８年９月号．(別册)， 機械翻訳，PP 128-134.