the procedure of measuring quality and progress of system tests - an outline

J. Ritzke SNI AP 344 MUC
13/7/92

Facts:

the system functionality and quality of METAL language pair translation can be evaluated as follows:

1. benchmark-texts which have to be corrected step for step and be tested again

2. additional benchmark texts in order to evaluate the expansion of functionality which also have to be corrected and tested successively

in principle the functionality and quality of these components is evaluated in two ways:

1. grammar evaluation
   i. e. correctness and coverage of structure description as analysed by grammar and procedures (so-called LingWare)

2. lexicon evaluation
   i. e. correctness and coverage of the vocabulary

Test procedure to measure functionality:

a. 1/ one text is tested several times and 2/ other texts are added

b. 1/ structures and grammar and 2/ lexicon

we propose the following two level procedure to calculate and present the improvement of system's quality:

1 level: one and the same benchmarktext:

a. test procedure :

i. preliminary remark:
the available test tools (diff benchmark tools and CSL-tools) which are usually applied in METAL can be used to evaluate and compare test results:
here only the differences, i.e. improvement or disimprovement of the current test, are printed
identical test units which have to be involed in calculating the value of the quality changes (see below: factor I) are automatically indicated by the decreasing number of translation units in the diff-file

ii. starting: first translation of the text

iii. 1st step: first correction of the test results by improving structural description and lexicon

iv. 2nd step: second translation

v. 3rd step up to step n: second correction, third translation, third correction ... and so on

b. test evaluation :

the factor of quality changes (QV) is calculated with respect to the following criteria which will be set manually by the test person after the examination of the current test results as to the dissimilarity to the former translation :

i. B_S = structural improvement (= result of lingware corrections)

ii. B_L = lexical improvement (= result of lexical corrections and additions)

iii. I = absolute identity

iv. G = similarity (= equivalence of form and content)

v. S_S = structural disimprovement (= potential side-effects of lingware corrections)

vi. S_L = lexical deterioration (= potential side-effects of lexical corrections)

c. calculation of quality changes :

==> FORMULA : $QV = 2*B_S^2 + B_L^2 + I + 2*G - 2*S_S^2 - S_L^2$

evaluation factors :

i. grammatical corrections are more complicated and more tim-consuming than lexical corrections and additions
that is why the evaluation of these phenomena is different; cf.:
$2 * B_S^2$ and B_L^2
and $2 * S_S^2$ and S_L^2 respectively

ii. the evaluation of absolute indentity is neutral:
I

iii. similarity -i. e. equivalence with regard to form and content- is evaluated by a lower factor:
2 * G

d. some notes regarding QV :

i. QV refers to one sentence only (see below, ?)

ii. a satisfactory quality of the system corresponds to the predifined QV threshold value

iii. the results of test evaluation achieving the threshold value can be described by a curve:

1. ordinate: QV of the current test

2. abscissa: number of test phases (also correction phases ~~correction~~)

iv. the new QV is added to the former QV; it will be subtracted in case of deterioration (= negative QV value)

v. starting values :
QV = 0 , number of tests = 0;
status : the first translation is finished, comparative result values are not yet available

2 Level: additionnal benchmarktexts :

a. Test procedure :

   i. starting: one and the same benchmark text has passed the given threshold value i. e. a satisfactory system functionality is reached with regard to this text (end of first level)

   ii. 1st step: a new benchmark text is composed by adding to the old one another text covering further -amplified- system functionality

   iii. 2nd step: the new benchmark text consisting of the old one and an additional text is treated in the same manner as the old text (cf. first level) until the given threshold is passed

   iv. 3rd step up to step n: other texts are added in a similar manner and further benchmark tests whith further text additions are carried out successively until the given number of texts is translated.

b. test evaluation :

   i. the old text and the new one are tested simultaneously in order to avoid side-effects in the translation quality of the old text while the new one is tested

   ii. the current text which consists of an old and a new text is evaluated by the same test procedure as described above (first level).

   iii. a satisfactory quality is reached at the second level when the added text needs as few new test and correction periods as possible
   (Ideal = 0: the system is able to translate every text in a satisfactory manner (QV-threshold passed))
   (realistic proposition: threshold value : one correction phase at the second level)

   iv. also at the second level, the quality of the system corresponding to the results of the test phases can be described by a curve with the following values :

      1. ordinate: number of the correction phases of the current text (consisting of an old text and an added new one)

      2. abscissa: number of the added texts

   v. the quality improvement can be measured by examining the gradient of the curve

3. **outstanding questions:**

   a. does the QV refer to a sentence, a translation unit or to every current phrase inside a sentence ?

   b. should further factors like the evaluation of so-called phrasals be involved in the calculation of the QV (their number and their elimination: 3 * P) ?

   c. appropriate text corpora have to be made up in order to augment the system quality in a consistent way (cf. paper Dr. Thurmair, CSL-proposition)

   d. the same text corpora should be used for several language pairs with a common source or target language

4. **Tools :**

   a. to be developed:

      i. evaluation tools with user-friendly interfaces including windows, menus etc. (for the whole test process of level one and two)

      ii. automatic generation of statistical curves by this toolbox (in this way, quality can be defined in an objective manner)

   b. available:

      i. translation evaluation tools of CSL (cf. paper F. Deprez : Translation Evaluation Tools ("Sysiphos") (1992))

      ii. "diff"-lisp-functions at MUC (compare-translation etc.)