Towards a MT Evaluation Methodology

Adriane Rinsche

The Language Technology Centre
E-mail: XE_S412@uk.ac.Kingston

This paper is based on extensive studies of MT evaluation methods and practical comparative evaluation experience testing major MT systems such as Systran, Metal, Logos, Ariane. Evaluation studies were carried out within the framework of a Ph.D., on behalf of the EC Commission, and the London based computer consultancy OVUM. The paper begins with some general observations on the state of the art of NLP and MT evaluation, followed by a presentation of evaluation methodologies for comparing the quality of the linguistic performance of either subsequent versions of one machine translation system (vertical evaluation) or of several translation systems (horizontal evaluation). Merits and shortcomings of a rudimentary corpus based vertical evaluation methodology are briefly discussed. The horizontal evaluation method introduced consists of a combination of test suites and text samples. Translation quality is measured quantitatively by counting error frequencies. Some recommendations are also made regarding comparative linguistic performance testing of translation workbenches, particularly with regard to linguistic versus statistical fuzzy matching efficiency.

I. The present situation

During the last thirty years, numerous attempts have been made to develop quality criteria for NLP systems. However, up to now, no generally accepted evaluation procedure has emerged. Developers, researchers, sponsors and users are therefore forced to create their own set of evaluation criteria and techniques or to use methods developed for a specific NLP system or component which cannot be readily transferred to another system or application.

II. European Evaluation initiatives

In order to overcome this unsatisfactory situation, a number of European initiatives have recently been taken.

1. LISA is the "Localization Industry Standards Association", a non-profit organization with the objective of proposing methodologies and standards to ensure that the creation of multilingual software, technical documentation, and

the multimedia production process is globally achieved to the highest possible quality levels meeting the needs of the end users. LISA came into existence approximately two years ago.

2. EAGLES is an "Expert Advisory Group on Language Engineering Standards", launched within the framework of the CEC's Linguistic Research and Engineering (LRE) Programme. It is conceived as a coordinated set of European expert groups, each with a mandate to work towards common specifications and guidelines in various fields of interest to the NL and speech groups.

3. EAMT, the European Association for Machine Translation, considers evaluation as one of the key issues to be addressed.

4. Within the German MT user group, a large evaluation group consisting of actual and prospective MT users has been set up on an informal basis.

III. Relevance of evaluation

Such initiatives are indicative of the relevance of the issue. Evaluation is a key component in every technology. It is necessary to assess the general performance of systems. Performance can be measured for different development stages of the same system. This type of comparative evaluation will be called "vertical". Alternatively, the quality of several systems can be compared at a specific point of time. This second type of comparative evaluation will be called "horizontal".

Vertical evaluation is relevant to NLP product developers and service suppliers.

Horizontal evaluation is relevant to both developers and end users. To developers in order to determine the relative quality of their systems as compared to their competitors: to prospective end users in order make an informed judgement of what product(s) might best meet their requirements.

IV. Evaluation Constituents

NLP quality assessment is required with respect to the following parameters:

- software design
- user friendliness
- linguistic performance
- organizational implications
- cost effectiveness

Evaluation is always application-specific when carried out on behalf of prospective users.

In a global evaluation context (cf. DARPA) subject matters should be revealed beforehand in order to allow system developers to prepare for the lexical requirements of such evaluation.

V. Methodology Requirements

A given methodology should lead towards international standardized procedures of NLP benchmarking.

The methodology must be pragmatic and transparent to (prospective) end users.

Quality judgments of NLP systems should be achieved on a quantitative basis.

This should happen within a reasonable time scale at reasonable cost.

Evaluation results should have an impact on future more user oriented software development.

VI. Methodology Overview: Linguistic performance evaluation of MT systems

The MT evaluation methodology suggested below was preceded by a review of the relevant evaluation methods used in the past.

1. Vertical evaluation

A rudimentary measure of translation quality improvements of updated MT system versions as used with the Russian English Systran version in the United States is based on a translation memory like corpus of source and target text material. Preliminary tests have been carried out at the European Commission for all 16 language pairs (according to recent personal communication with J.M. Leick, DG XIII, CEC).

With a simple programme, all sentences translated by the current system version (target version n+1) are compared with all sentences translated by the previous system version (target version n). Sentences with identical translations are eliminated, all differing translation versions (between roughly 3% and 12% of the corpus) are printed in a context of 2 or 3 sentences and presented to three classes of evaluators: end users, system developers, and translators. Evaluation is rudimentary, because the only information requested from evaluators is whether the quality of the new translation is better, the same/similar or worse compared to the previous one. System developers may subsequently evaluate the translation results in more detail and use the information for further system updates.

This fairly crude evaluation procedure has been chosen, because a text corpus of approximately 10 000 words can thus be evaluated within a week. The percentage of improved versus deteriorated sentences is a crude quality indicator and could be improved by asking the evaluators to give reasons for their judgement, thus structuring their line of thought, by ticking one or more of the additional options:

Translation is better, because

- the number of lexical errors is reduced
- the number of syntactic errors is reduced
- the number of semantic errors is reduced

Translation is worse, because

- the number of lexical errors has increased
- the number of syntactic errors has increased
- the number of semantic errors has increased

(indicate difference: number of errors in target version n - number of errors in target version n+1 per category)

Translation is of similar quality with minor modifications

- at the lexical level
- at the syntactic level

(indicate number of changes)

The number and possible refinement of linguistic criteria (cf. below, 5.2.) used in a thus combined quantitative/linguistic approach would be subject to experimenting in order to arrive at a highly pragmatic and informative method.

The current SYSTRAN evaluation approach is based on the assumption that a new system version will produce a large number of identical translations.

This approach cannot be readily transferred to other systems if the percentage of differing target version n and target version n+1 translations is too high, because it would involve tremendous human effort and investment to evaluate large corpora without prior elimination of a substantial part of the corpus. This situation may occur if

(a) large scale system modifications have been carried out or

(b) a system is designed to generate different target versions of the same source sentence and does so frequently

(c) the corpus consists of highly complex embedded sentence patterns

Under such circumstances, this method cannot be used for vertical evaluation. It is unsuitable in any case for horizontal evaluation. When different translation systems are compared, the number of identically translated sentences must be expected to be minimal, because

(a) each source sentence allows for a range of acceptable translation equivalents and

(b) for an even larger range of error sources depending on the respective system and dictionary design.

It is therefore necessary to adopt a different approach for horizontal evaluation.

2. Horizontal evaluation

Initial Product Overview

For each MT system and workbench to be compared, systems specifications and characteristics should be described in the form of a grid to allow for an initial product overview.

The following information should at least be covered:

System name	
System developer	
Languages Covered	
Subject Areas Covered	
MT System Type	(direct/transfer/ interlingua)
Price	
Specification	(Hardware, Operating system, Processor, Ram, Hard disk space, Monitor, Other)
Editor	(Standard/Application specific)
Processing	(Interactive/Batch)
Network Compatibility	
Record Formats	
Filters	
Format Protection	(Source/Target)
Terminology Organization	(Monolingual/Bilingual/Multilingual/Reversible)
Translation Memory	(Identical/Linguistic/Statistical Fuzzy Match)
Database Management Facilities	
Grammar design	
Semantics	

Linguistic Performance Testing

The important specific feature to be tested in the case of Workbench Products is efficiency of fuzzy matching, statistical or linguistic. A detailed procedure of comparing fuzzy matching design is yet to be developed. The approach must be corpus based and count at least the number and relevance of successful fuzzy matches. It would be a very interesting research task to design a suitable evaluation framework to compare the efficiency of linguistic and statistical fuzzy matching.

For machine translation systems, a more refined technique has been developed:

In each testing cycle two types of test material should be used:

1. Test Suites
2. Text Samples

A test suite is a set of sentences designed to test the range of grammatical features covered by a system. It should be global and context neutral as far as possible. In a comparative context it thus allows to judge the relative grammatical sophistication of systems.

A text sample is used to test the suitability of a system for use in a specific application domain.

The tests should be run in the following order:

Step I: Text samples only: Raw translation without prior
dictionary update
Measure translation time

Step II: Introduction of all necessary dictionary entries
Measure dictionary update time

Step III: Text samples: retranslate
Test Suite : translate
Measure translation time

Step IV: Evaluation of raw translations obtained in steps I and III by
- counting the number of error-free sentences
- counting the number of errors occurring
in test suite and text material separately

Step V: Interpretation and comparison of error statistics

Error Analysis

Both types of test material are linguistically analyzed. All errors are assigned to a number of categories in order to determine system specific error clustering in any of the following categories.

Category I : Lexical errors

Lexical errors are assigned where lexical items are not or incorrectly translated. As a subclass, verb errors are counted separately, because verb error frequencies are highly indicative of linguistic quality due to probable syntactic consequences resulting from faulty verb translation. The following categories thus emerge:

Lexical item - no translation
Lexical item - wrong translation

Verb - no translation
Verb - wrong translation

Category II : Morphological errors

Morphological errors mainly refer to subject verb agreement:

Word formation error

Category III: Syntactic errors

This error category refers to sentence generation problems without semantic effects. The meaning of the sentence, phrase, or clause is still recognizable. The following subcategories are distinguished:

Sentence structure - wrong
Verb phrase - wrong
Noun phrase - wrong
Prepositional phrase - wrong
Subordinate clause - wrong

Incorrect sentence structure refers to a sentence with an overall messy syntactic structure, with the sentence meaning still recognizable. A complex "hyper"sentence consisting of several coordinated "basic" sentences is divided into these substructures, each of which counts as a sentence.

Verb, noun, and prepositional phrase syntactic errors refer to faulty target phrase generation. Syntactic errors in coordinated complex noun and verb phrases are counted separately.

The same principles apply to subordinate clause evaluation.

Category IV: Semantic errors

Semantic errors may affect either the meaning of the whole sentence, or the meaning of noun, verb, or prepositional phrase. Further categories comprise reference errors and idiomatic expressions:

Sentence meaning	- wrong
Meaning of noun phrase	- wrong
Meaning of verb phrase	- wrong
Idiomatic expression	- wrong
Reference error	

Coordinated sentences or phrases are counted and analyzed separately as in Category III.

Category V: Source text error

Source text errors should not appear. In case they are found in the evaluation, they are due to

- typing mistakes in the test material or occurring in the course of transferring the text material into system processable form

- source text ambiguities which cannot be resolved by the system.

They should be counted separately in order to do the systems justice.

Quality Assessments and Interpretation

The following quality assessments can be derived from the data achieved:

1. Quality of "blind" translation (text samples only)

   based on error counts - quantitative

2. Quality of dictionary updating facilities by analysing

(a) user friendliness, linguistic sophistication and suitability of dictionary design
- descriptive

(b) time required for update
- quantitative

3. Quality of translation after dictionary update for

- test suites and
- text samples, respectively

based on error counts, with very interesting different, test type and system specific results in each type of test material, leading to a quantitative

- grammatical performance profile
- three dimensional suitability profile as a function of system/language pair/subject domain

4. Overall quality per system tested:
interpretation of benchmark results - descriptive

5. Systems Comparison - descriptive and quantitative

Previous tests have shown that error clustering varies considerably depending on the test and system type. While test suite error frequencies give evidence of a system's grammatical coverage in general, text sample error counts are indicative of the suitability of a subset of grammatical features for a given application.

In a small number of cases a certain degree of subjectivity may be retained in the process of assigning errors to categories, because MT output analysis without "glass box" information may lead to problems with determining error causes correctly. MT quality comparisons consistently based on analysing surface phenomena via error analysis are, however, the most objective way of linguistic performance measurement. Comparisons of error frequencies across systems are based on quantitative results and are therefore a suitable and reliable basis for comparative quality statements. This procedure is the only objective and quantifiable means of direct translation quality measurement. The fact that large software companies such as DIGITAL measure human translation quality in terms of error frequencies further confirms the approach adopted. It will be natural to MT users to adopt the same strategy for MT output evaluation.

VII. Conclusion

Evaluation of natural language processing software is a young and immature discipline. A lot of work must be done before satisfactory methods and internationally accepted standards will be available. The very nature of natural language precludes fully objective and formal evaluation methods. No matter how quantitative and automatic a procedure will be, a certain element of subjectivity will remain. It is necessary to minimize this element and the amount of human effort involved in evaluation in order to arrive at practicable and financially viable solutions. The above described procedures will hopefully stimulate further discussion.

References

Bourbeau, Laurent. 1990. Elaboration et mise au point d'une méthodologie d'évaluation linguistique de systèmes de traductions assisté par ordinateur. Rapport final. Secretariat d'Etat du Canada.

Falkedal, Kirsten. A Practical Guide to the Evaluation of Machine Translation Systems. ISSCO: Interim Report to Suissetra.

Rinsche, Adriane. 1993. Evaluationsverfahren für maschinelle Übersetzungssysteme. Zur Methodik und experimentellen Praxis. Ph.D. Thesis. Kommission der Europäischen Gemeinschaften. Informationsmanagement. EUR 14766DE. ISSN 1018-5593.

Rinsche, Adriane. 1991/9. Towards a System of Benchmarking MT Systems. Report for EC Commission. October 10, 1991.