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

Comparative MT Performance Evaluation: an empirical study

Adriane Rinsche
University of Bonn

The subsequently described study was carried out within the framework of a PH.D. project on machine translation evaluation. A methodology was developed, designed to measure MT systems performance.

Of course, there is a great variety of systems available on the market. Only a few could be evaluated within the given framework. The choice was determined by several parameters:

1 Language pairs

German to English was chosen as the most interesting language pair – German as a source language due to the German nationality of the author and English as a target language due to the fact that English is one of the most important target languages, not only with respect to MT, but regarding NLP in general.

2 PC based versus mainframe based MT systems

PC based systems were ruled out, because they are so cheap that a trial and error procedure presents no risks to a (prospective) user.

The main frame systems covering German to English are METAL, LOGOS, and SYSTRAN. The tests with Logos and Metal were carried out at GMD, Bonn, and with Systran at the EC, Luxembourg, between August 1990 and August 1991.

3 General objectives of the evaluation

The study is aimed at measuring MT quality from a user's point of view.

Although extralinguistic aspects such as the possible integration of MT hardware and software into an existing DP environment and how to fit MT into an existing organization – including staff requirements, office modifications, ergonomic aspects etc. – play an important role in any overall comparative evaluation, these aspects were ruled out for the purposes of the study. Since the evaluation was not carried out within the framework of a specific application, it was possible to concentrate on the relative linguistic performance of the systems.

From the variety of evaluation procedures described in the relevant literature, error analysis was chosen as the most objective approach. The considerations leading to this decision are described in more detail in A. Rinsche (Ph.D. thesis, forthcoming 1992).

Although microevaluation (van Slype 1982:54), that is, improvability of performance, is a very important issue, the tentative "black box" evaluation carried out provides only very fragmentary insight into any MT linguistic and software design. Meaningful conclusions on improvability can only be achieved by close cooperation with developers within a highly complex evaluation cycle. Such procedures, described by Margaret King (1990:214ff) are far too time consuming, complex and expensive to support a decision making process within a limited period of time.

4 Choice of test material

Evaluation of authentic text samples is user oriented and realistic. Real language is evaluated in a real context. Texts are chosen according to their relevance to a planned or existing application. As far as representativeness is concerned, there is much controversial discussion but no resolution as to what sample size might be adequate. For pragmatic purposes, a sample size of 100 to 150 sentences, depending on the respective syntactic complexity chosen, might suffice.

Test suites, on the other hand, are designed to test linguistic phenomena systematically. They are linguistic artefacts, that is, sets of contextually unrelated sentences. The linguistic context required to test a specific phenomenon within the minimal context of a sentence influences the outcome. It is difficult to isolate this structure to be tested without interference of other contextual phenomena. The information gathered is useful to developers but unrealistic, because it provides no information about the usability of a given MT product in a prospective application environment. Test suites become too complicated if globally used. Importance of context, which may supply information to solve translation problems is not accounted for.

An evaluation based on a combination of both test suite and text sample material seems to be ideal.

5 Objective of pilot study

The study is aimed at developing an easy and practical test exposing some characteristics, strengths and weaknesses of translation systems in a realistic time scale at a realistic price, because hardly anyone can afford to wait for a year or to invest half the price of the main frame system into a preliminary evaluation.

An attempt is made to measure the capability of systems to deal with general purpose texts with respect to

- lexical coverage
- syntactic analysis/synthesis
- semantics - meaning preservation on word and sentence level.

The reason for choosing a general purpose text is that the evaluation is not aimed at a specific application. It was, however, considered to be important to gather information about machine translation quality of general purpose texts, because it is easier to become more specialized in a given application rather than to generalize from a specific domain.

6 Source text analysis

At the lexical level, the text sample chosen contains "general" vocabulary plus a small amount of DP terminology; there are no sophisticated verbs, adjectives, adverbs, or idioms; lexical coverage is tested only to the degree that the lexicon contains sufficient lexical items and differentiation categories to choose the right word in the right context.

There is a fairly high complexity of syntactic structures; average sentence length is 12.5 words. Nagao suggests a sentence length limit of 30 words for MT systems; longer sentences, he believes, cannot be handled properly by existing MT systems; this limit is reached twice. No attempt is made to include any pragmatic considerations such as discourse analysis, since none of the systems is designed to cope with phenomena above sentence level.

The text sample is believed to be representative of any general purpose text, although no statistical proof has been sought. Some linguistic restrictions are imposed by the text type and sample chosen. They are explained below (cf. "Explanation of error categories").

7 Evaluation strategy

7.1 Sample size

The overall sample size consists of four data sets containing at least 200 sentences. After analysing the first, biggest data set using the categories emerging from the work on the three translations significant differences in the performance of the systems became obvious. Numerical proof of impressionistic quality judgement was achieved.

7.2 Analysis of raw translations

The analysis of translations without permitting terminology entries is meaningful in the case of the sample chosen, because terminology problems are fairly minor. Lack of vocabulary does not seem to impair the results significantly. No or wrong translations occurred only in

- 67 words (Metal)
- 53 words (Logos)
- 57 words (Systran)

Had terminology entries been admitted, the comparability of systems would have suffered due to differing encoding facilities.

8 Explanation of Error Categories

The error categories were chosen during the course of the error analysis of the texts. In addition, Flickinger's comprehensive test suite for the English language was consulted (Flickinger, 1987). Only few of the linguistic phenomena described were used for this study, because Flickinger's apparatus was not devised for translation problems in particular, but has a more general orientation towards analysing NLP systems. With the source language German, however, quite specific transfer problems arise.

The second restriction arises from the fact that one specific text type was selected for the analysis. The sample was chosen from the descriptive text category. Descriptive texts display only a section of possible structures; imperatives and questions (except for one), subjunctives and indirect speech are omitted from this sublanguage altogether.

Third, the error typology had to be fairly rough in order to allow for clustering of typical characteristics. With too many, too specific or too detailed categories three individual evaluations would emerge which might not allow for any comparative statements.

The error categories chosen are ordered according to the traditional linguistic division into lexical, syntactic, and semantic errors.

Morphology as one further decisive linguistic discipline is ruled out for the purposes of the evaluation, because English as the target language chosen is morphologically poor. Word formation problems do not seriously affect translation quality as far as the three systems tested are concerned. Compound formation is accounted for at the lexical level; inflectional morphology and morphosyntactic problems are dealt with at the syntactic level.

Translations of 73 sentences carried out by each of the three translation systems were analysed. A sentence is defined as a sequence of lexical units limited by a full stop (.) or semicolon (;) or question mark (?) (in one case). Complex sentences consisting of a sequence of main and/or subordinate clause(s) were counted as one sentence. Error codes 80 and 90 (see section 8.4 below) referring to syntactic structure and sentence meaning may be assigned to one complex sentence more than once depending on the number of errors in each substructure. For example, sentence no 30, translated into English by Metal as:

- "It was introduced in 1889 and found in hochentwickelter Form Verwendung, until the electronic computer was introduced in the fifties and far circulation found it."

is encoded as follows:

- 30 ok /31 11 11 ,ok /80

The first subsentence within sentence no. 30 is error free (ok — "it was introduced in 1889"); in the second subsentence one adjective and two nouns remain untranslated; part three again is error free (ok = "until the electronic computer was introduced in the fifties"), whereas the last substructure receives error code 80 for erroneous syntactic structure. Although Metal recognizes the subject common to both subsentences and adds "it" during the source analysis phase, yet the system is unable to reorganise the structure correctly in English. The sentence meaning is still recognizable, but the sentence is neither idiomatic nor syntactically correctly organized.

8.1 The Dictionary

The dictionary is a valuable component of any translation system. For buyers and developers around 50% of the overall investment goes into the dictionary. According to Vasconcellos (1991) minimum dictionary size should be at least 25,000 entries for non-specialized dictionaries. This minimum requirement is by far exceeded in all three systems, and this fact is mirrored in the small

proportion of untranslated vocabulary.

The contextually adequate translation of lexical items depends on

1. correct encoding
2. the degree of semantic subcategorization
3. dictionary organization: number and interconnection of specialized subdictionaries.

Contextually adequate homograph disambiguation is one important MT dictionary quality feature, particularly where future extensions are planned. In the test Systran was most reliable as far as general vocabulary use was concerned, but quite error prone with respect to the small amount of DP terminology included in the sample. However, once a system is introduced, a more specialized application can more easily be introduced than a more generalized approach.

Some examples of incorrectly translated lexical items are given below. In several cases disambiguation will only be possible by introducing fine semantic distinctions which are probably beyond the scope of the mainly syntax based present MT linguistic design.

German source word	translated word	correct translation
Nouns		
Abbau	shut-down	reduction
Abbau	dismantling	reduction
Verkleinerung	reduction	minimization
Form	mold	form
Menge	set	quantity
Bedeutung	meaning	importance
Zeit	moment	time
Band	volume	tape
Verbs		
einstellen	adjust	employ
einsetzen	insert	use
erledigen	finish	deal with, settle
entstehen	result	develop
einführen	import	introduce
verlassen	desert	leave

Error counts are based on tokens rather than types, that is, each occurrence of each lexical error is counted, no matter how often the same error is repeated.

8.1.1 The "no translation" error category

Error types subsumed under the above label expose weaknesses in lexical coverage.

The small number of untranslated vocabulary in the sample confirms the low degree of difficulty of the sample text and the suitability of the sample. If too much vocabulary is unknown to the dictionary, syntactic analysis and interpretation of whole sentences becomes unreliable or even impossible. The fact that, for instance, in the case of the translation carried out by Systran, only 7 lexical items remained untranslated as opposed to 20 untranslated words in the case of the second version of Metal, may lead to slightly worse results in other error categories for Metal as well.

The error type "no translation" can in general be used to evaluate an MT system only if text samples containing mainly general vocabulary is used and if raw translations without dictionary updates are carried out. As soon as specialist vocabulary is contained in a text sample the suitability of the MT dictionary must be examined beforehand. If no additions are made although relevant information is missing, the evaluation may fail to lead to reliable results.

The evaluation presented in this study is based on raw translations because after post-editing using different staff and strategies the machine translation results as such are not comparable any more.

8.1.2 The "wrong translation" error category

Errors pertaining to this category expose weaknesses in lexical coverage and/or semantic subcategorization. Again, performance is rather satisfactory in all three systems, with slight gradual differences.

Particularly significant error frequencies arise in the case of verbs and prepositions.

The amazingly high number of verb errors in both versions of Metal may partly be due to that system's smaller dictionary size or encoding of specialized verb meanings.

Prepositions are another important source of error in German to English translation. The fact that all three systems reacted similarly error prone leads to the tentative conclusion that this seems to be a global translation problem for the respective language pair rather than a source of error due to more or less elegant linguistic design of a specific MT system. Systran displays specific weaknesses regarding the correct elimination or addition of

prepositions with altogether 11 occurrences as compared to only 1 in each of the two other systems. This error type obviously emerges in the synthesis phase.

The different use of articles in German and English is mirrored in the frequently erroneous elimination or addition in the target language. Logos seems to be most error prone, followed by Systran and Metal. It is interesting to observe that in Metal error frequencies deteriorate in the second version. Since the wrong placement of articles may depend on other, co-occurring errors and since in general sentence meanings are not affected by erroneous article placement in German to English translations, future pragmatically oriented evaluations regarding this language pair may do without this error type altogether.

Wrong translations of conjunctions do not lead to any distinctive quality differences. Due to the small variance regarding meanings and translation equivalents this word class is of limited difficulty and can be represented fairly easily by a MT dictionary.

Pronouns are in general correctly translated. Reference errors are quite rare.

Nearly all adjectives are translated and except for one are correctly translated into the target language. None of the systems tested was particularly error prone in this respect because the adjectives used in the sample are stylistically fairly neutral.

Systran is most successful as far as adverb and adverbial phrase translation is concerned. Under this category the choice of a wrong adverb is encoded but also a word formation error, namely missing adverbial suffix "-ly".

The very low number of incorrectly and untranslated compounds is partially due to the fact that the sample chosen is linguistically not too complicated. Some of the wrong compound translations cannot be easily corrected by further subcategorization or by introducing extra rules. Among these the translation of "Absatzlage" as "paragraph position" rather than "sales situation" and "Gehaltszettel" as "content label" rather than "salary note" may serve as examples. The translation of "Rechenautomat" as "rake automat" rather than "calculating machine" might be more easily avoidable, particularly because the compound is defined in the same sentence.

8.2 Syntax

The text sample contains many different syntactic structure elements. Due to the text type (descriptive) some structures are not represented. Some sentences are simple, others quite complex with deep embeddings. Average sentence length is 12.5 words. Evaluation of syntactic features is restricted

to just a few features. Error types are subclassified as follows:

Erroneous syntactic organization

- of parts of sentences such as verb and noun phrase,
- of subordinate sentences, and
- of main clauses with a completely destroyed syntactic structure.

Complex sentences consisting of more than one main clause are analyzed separately.

8.2.1 The "Sentence structure wrong" error type

The fairly high error frequency in this category demonstrates the difficulties arising during analysis, transfer and synthesis of complex sentences. The corresponding error code is assigned when sentence structure is destroyed but sentence meaning is preserved. If sentence structure errors lead to sentence meaning distortions a semantic error is assigned instead.

8.2.2 The "Verb phrase - wrong construction" error type

As might be expected in the descriptive text type the potential structural range of verb phrase complexity is exploited only to a limited degree. Present, perfect and imperfect tenses are used mainly in the active voice. There are few examples of the passive voice. Modal verbs occur rarely and in a low degree of complexity as far as their combination with other verb phrase elements is concerned. Errors arising due to wrong adverb placement are subsumed under this category as well.

Logos is slightly more successful than Metal as far as verb phrase construction is concerned. The high frequency of verb phrase errors in Systran may be due to the fact that Systran is far more successful in retaining sentence meanings. It is only logical that with a higher number of preserved sentence meanings more errors might occur at lower levels of complexity.

8.2.3 The "Noun phrase - wrong construction" error type

This error category is used when elements within the noun phrase are in the wrong place or when case, number or gender are incorrectly allocated.

8.2.4 The “Subordinate clause - wrong construction” error type

The subordinate clause is defined as a dependent sentence. The most frequently occurring subordinate clause in the sample is the relative clause. Problems arise in the distinction between identical German definite articles and relative pronouns “der, die, das”. The number of subordinate clauses of this type is low in the text sample chosen and therefore does not provide a major source of error.

8.3 Semantics

Semantic errors are restricted to very few categories since the whole range of word meanings is already covered in the different lexicology error categories (cf. Dictionary above).

8.3.1 The “Sentence meaning not understandable” error type

This error type is the most serious of all. When this error code is assigned, no other, more detailed codes at lower levels of complexity are used, because the destroyed sentence meaning does not allow error cause identification at lower levels. Systran performs outstandingly well at this level and must therefore be classified as the most mature system, if a conclusion like that is derivable from a limited test of this scale at all.

8.3.2 The “Negation in the wrong place” error type

This error type is assigned when erroneous negation placement results in sentence or part of sentence meaning distortion. When only syntactic arrangement is concerned, the corresponding syntactic error is assigned at the level of sentence structure, verb or noun phrase error.

8.3.3 The “Idiomatic expression wrong” error type

This source of error is difficult to resolve by machine translation systems at the present development level except by encoding the respective phrases separately. Automatic recognition and target language specific translation of syntactically more or less complex idiomatic expressions cannot be expected. Literal renderings of idiomatic phrases may lead to meaning distortions of parts of sentences. If the literal translation results in a clumsy but recognizable meaning no error code is allocated. The low frequency of errors is due to the text type and relative stylistic simplicity.

8.4 Error Categories - Distribution in the Translations

		1	2	1	2	
LEXICON						
11	Noun - no translation	9	8	3		1
12	Noun - wrong translation	6	6	9	7	9
21	Verb - no translation	3	3			
22	Verb - wrong translation	18	18	5	6	6
31	Adjective - no translation	4	4			3
32	Adjective - wrong translation	1			1	
33	Adverb(ial phrase) no translation	3	3	2	2	
34	Adverb(ial phrase) wrong transl.	7	5	4	1	1
41	Preposition - no translation	1	1	2	1	
42	Preposition - wrong translation	9	9	11	9	9
43	Preposition - not added	2	2	1	1	7
44	Preposition - incorrectly inserted			1		4
51	Pronoun - no translation					
52	Pronoun - wrong translation	3	7	3	1	5
61	Compound - no translation	1	1		2	3
62	Compound - wrong translation	7	8	9	9	6
75	Article - added by mistake	4	5	6	6	3
76	Article - not added though needed	1	2	2	2	3
SYNTAX						
80	Sentence structure wrong	12	8	14	10	15
81	Verb phrase - wrong construction	6	7	4	5	10
82	Passive not or incorrectly constr.	2	3		3	
83	Tense/aspect - wrong formation	3	1	1	1	1
84	Modal/aux. phrase - wrong form.	5	4	3	1	4
85	Noun phrase - wrong construction	6	5	4	3	3
86	Subordinate clause - wrong constr.	2	2	2	3	3
89	Singular/Plural agreement	4	3	7	3	6
SEMANTICS						
90	Sentence meaning not understand.	10	7	8	6	1
92	Negation in the wrong place			2	1	
93	Idiomatic expression wrong	2		2	2	3

9 Summary

Evaluation is a tricky business. It is very difficult to set up criteria applicable to a wide variety of MT systems in the same manner in order to reliably compare MT quality.

Error counts are impaired by the restrictive insight into the “black box”. Simplifications cannot be avoided.

The results provide, however, interesting information about error frequencies in raw translations emerging from different releases of different MT systems regarding one language pair. The study does not claim to be representative. It is restricted to a sample taken from one of a variety of several possible text types. In a “real life” industrial, political or administrative context different requirements may lead to using different samples from different text types leading to different results. Other target languages may require a modified error type configuration. Objectivity may be impaired by dictionary size variations and specific design features.

The test is, however, simple and easy to administer. More recent research extending the preliminary case study to one other language pair, to further MT systems, text samples and test suites will be described in the Ph.D. thesis mentioned (A. Rinsche, forthcoming 1992).

Literature

Flickinger, D. et al. 1987. Toward Evaluation of NLP Systems. Stanford Linguistics Institute.

King, M, Falkedal, K. 1990. “Using Test Suites in Evaluation of Machine Translation Systems”. Proceedings. Coling '90. Helsinki.

Rinsche, A. Forthcoming 1992. Vergleichende Performanzevaluation Maschineller Übersetzungssysteme. Ph. D. Thesis.

Van Slype, George. 1982. Etude et Mise au Point de Critères d’Evaluation techno-économique d’un Procède de Traduction Automatique. Bruxelles.

Vasconcellos, M. 1991. “Perspectives on the Assessment of Machine-translated Output”. FIT Miscellany on Translation Criticism. Hrsg. Milan Hrala.