

Correspondence Measures for MT Evaluation

Lars Ahrenberg & Magnus Merkel

Department of Computer and Information Science
Linköping University
S-581 83 Linköping, Sweden
{lah,mme}@ida.liu.se

Abstract

The paper presents a descriptive model for measuring the salient traits and tendencies of a translation as compared with the source text. We present some results from applying the model to the texts of the Linköping Translation Corpus (LTC) that have been produced by different kinds of translation aids, and discuss its application to MT evaluation.

1. Introduction

In this paper we present a correspondence model and a set of measures that are of interest to linguistic evaluation of translations, including MT evaluation. As a point of departure we contend that the methodology of linguistic evaluation of machine translation systems must be set within the more general framework of evaluating translations, irrespective of the means and processes used to produce them.

A counter-argument to this view is that MT output, because of its deficiencies, serves different functions than translations produced by humans. A strong defender of this view is Sager (1993) who argues that MT output is best treated as a text type of its own “with a limited range of possible functions” (op.cit. p. 263) and that MT evaluation should focus on the appropriateness of output for the intended use. Of course there can be no denial that machine translation of today has limitations and that it is often used for other purposes than human translations, such as reading assistance of foreign language texts and cross-language information retrieval. Current developments in MT technology and MT use, however, make it more and more difficult to uphold a strict division between MT and HT. On the one hand, corpus-based methods such as statistical and example-based MT try to make use of human translation strategies as they are manifested in existing translations; thus it is of great interest to see how far these methods are able to succeed. On the other hand, MT is more and more used as a component in larger systems; for example in machine-aided human translation where a translation memory component, a term bank and various other tools are included. Again, with such a setup it is of interest to know how the end product compares to the translations produced by human translators without specific translation aids.

2. The Correspondence Model

The correspondence model is aimed at describing and quantifying structural and semantic relations between source text and translation; a common practice in descriptive translation studies (cf. e.g. van Leuven-Zwart, 1990). Thus, the goal is to provide a description of the

salient traits and tendencies of a translation as compared with the source text. The description is given by a set of quantitative measures that are calculated on the basis of a detailed analysis of structural and semantic shifts (Merkel, 1999; Ahrenberg & Merkel, 1997). For evaluation the method enables comparisons between different systems and different ways of organising a translation project. It also makes possible the characterisation of the translation norms that may exist for a given text type at a certain time (or in a certain organisation) and hence comparisons of a given translation with the norm. So far, we have applied the model to the translations in the Linköping Translation Corpus (LTC).

An overview of the Linköping Translation Corpus is given in Table 1 below. All bitexts in LTC are translations from English into Swedish. Some of the translations have been produced with the aid of translation tools, primarily translation memories, but one sub-corpus is machine translated.

The shifts that are recorded are optional shifts. Obligatory translation shifts (i.e. shifts that are necessary for grammatical well-formedness) are not categorised explicitly. The basic idea of the model is to establish and classify translation correspondences at the following ranks:

- sentences (as textual units, including headers)
- clauses
- clause constituents (incl. subject, main verb and various types of complements and adjuncts)
- phrase nuclei

It should be noted that the analysis does not consider every possible structural aspect of the translation. Changes concerning function words such as conjunctions, prepositions and articles, or the number and definiteness of nouns, are ignored in all cases except when they influence the classification of clause function or type of phrase. The semantic classification considers whether the two sentences of a pair have the same meaning, if there is more information on either the source or target side or whether there is another type of relationship, e.g. different or conflicting information.

Table 1. The Linköping Translation Corpus - an overview

	Text type	Title	No. of source words	No. of target words	Translation method
1	User's Guide	Microsoft Access User's Guide	179,631	157,302	Human (traditional)
2	User's Guide	Microsoft Excel User's Guide	141,381	127,436	Human (traditional)
3	User's Guide	IBM OS2 User's Guide and Installation Guide	127,499	99,853	Translation memory
4	User's Guide	IBM InfoWindows User's Guide	69,428	53,619	Translation memory
5	User's Guide	IBM Client Access for Windows User's Guide	21,321	16,752	Translation memory
6	Novel	Gordimer : A Guest of Honour	197,078	210,350	Human (traditional)
7	Novel	Bellow: To Jerusalem and Back	66,760	65,268	Human (traditional)
8	Dialog text	ATIS dialogues ¹	2,179	2,048	Automatic (MT)
Total			805,277	732,628	

Below the sentence rank, the model registers changes of various kinds. Correspondences that are not registered are assumed to be equivalent both structurally and semantically. A translation may involve non-1-1-mappings such as splits, convergences, additions, deletions and paraphrasing as well as shifts in rank, function, order, morphological properties and meaning. For a pair of corresponding segments at the sentence rank, all changes at the ranks below are registered and counted. These counts constitute the underlying data for further classification at the sentence rank and the calculation of indices that describe the translation as a whole in a certain respect.

The optional structural changes that can occur in the translations are analysed in the following categories and subcategories:

1. Changes related to the function and properties of clauses:
 - Voice shift (e.g., active > passive)
 - Sentence mood shift (e.g., imperative > declarative)
 - Shift of finiteness e.g., finitival > infinitival verb construction)
 - Level shift (e.g., main clause > subordinate clause, clause > phrase)
 - Function shift (e.g., temporal clause > conditional clause)
2. Changes related to the function and position of constituents:
 - Function shifts (e.g., manner adverbial > predicative)
 - Level shifts: (e.g., phrase > clause)
 - Transpositions (changes in order between constituents)
3. Changes related to the number of constituents:
 - Additions
 - Deletions
 - Divergences (one source constituent > two or more target constituents)
 - Convergences (two or more source constituents > one target constituent)
4. Paraphrases, which influence at least two constituents and cannot be split up into several smaller changes.

Changes of type (1)-(3) are regarded as *simple* while paraphrases are inherently *complex*. Note also that the simple changes can involve several constituents. A shift of sentence mood from imperative to declarative implies that a subject (implicit in the source) has been made explicit in the target. This leads to an addition of one constituent in the translation, but because this change can be seen as a necessary implication of the mood shift, this particular type of addition is not recorded separately.

Each sentence in the sample has been given an analysis in *translation segments*. A translation segment is defined as a nucleus, which is constituted by a content word, or a multi-unit term, and its accompanying functional words. Furthermore, main and subordinate clauses as well as a variety of phrases are also identified. Phrases are categorised based on their syntactic function in the clause. For each sentence pair the number of translation segments and different kinds of changes are recorded. The example,

SOURCE: |Twenty^I |people^{II} |gave^{III} |false^{IV} |testimony^V |.

TARGET: |Tjugo^A |personer^B | vittnade^C | falskt^D |.

is given a structural description which among other data contains the following information:

Table 2. Structural description including number of translation segments and changes from source to target

Translation segments, source:	5
Translation segments, target:	4
Convergence, (III+V->C):	1

2.1. Structural correspondence

Each sentence pair is attributed a structural correspondence feature as (i) *isomorphic*, (ii) *semi-isomorphic* or (iii) *heteromorphic*. An isomorphic relationship entails structural likeness, semi-isomorphic correspondence means slight structural differences and heteromorphic correspondence is used to signal major structural differences.

A translation is considered *semi-isomorphic* if it either (i) contains one simple change at the most, or (ii) contains at least seven translation segments and involves two simple changes at the most.

¹ The short text translated automatically was kindly provided by the Swedish Institute of Computer Science (SICS) in Stockholm and was translated with the SLT system (Agnäs et al. 1994).

A translation is considered to be *heteromorphic* if it contains either (i) one paraphrase, or (ii) at least three simple changes, or (iii) exactly two simple changes and less than eight translation segments in the source sentence(s).

The borderline between semi-isomorphic and heteromorphic translations is chosen arbitrarily, but it still serves the purpose of giving a rough estimation of how “free” or “paraphrastic” a translation is.

2.2. Semantic correspondence

The structural analysis of each translation pair is complemented with a comparison of the corresponding translation segments’ meanings. Cases of non-synonymy fall into three categories:

- (i) More specific nucleus,
- (ii) Less specific nucleus,
- (iii) Nucleus with a different meaning.

The classification of the semantic correspondence between source and target sentences is made in a similar fashion. Pairs of source and target sentences are classified as belonging to different categories depending on correspondences between their primary segments. In this classification four major categories are used:

- EQ:** Source and target sentences are regarded as having the same meaning;
- LSP:** The target sentence provides less information (is less specific) than the source;
- MSP:** The target sentence provides more information (is more specific) than the source;
- OTH:** Source and target sentences have some other relation to each other than the above.

Translations are considered to be EQ if no isomorphic translation alternative with closer semantic correspondence than the actual translation can be formulated.

It should be noted that the classification of semantic correspondence is based on the structural and semantic analysis of constituents and does not rely on the interpretation of the content in context. A translation pair is placed under the MSP category if the only changes in the translation are additions or the occurrence of more specific lexical items. Consequently, a pair is categorised as LSP if the only changes are deletions or the occurrence of less specific lexical items in the target. If there is a mixture of such changes (both MSP and LSP contributing changes), or if simple functional changes give rise to semantic effects, the translation pair is categorised as OTH. A paraphrase is often meaning preserving, but in exceptional cases it can cause a change of meaning and lead to a different categorisation than EQ.

3. Measures

At this point the first measure for correspondence can be given. A simple measure that indicates the proportion of isomorphic translations is defined as:

$$\text{Isomorphy Index} = I/N$$

where I is the number of isomorphic sentence pairs and N is the number of sentence pairs in the sample.

More fine-grained measures of structural correspondence can of course also be obtained. The following is one way to quantify semantic change:

$$\text{Structural Change Index} = (H+SI/2)/(I+SI+H)$$

where I , as above, is the number of isomorphic pairs, SI the number of semi-isomorphic pairs and H is the number of heteromorphic pairs. The measure is designed so that it returns the value 1 if all sentence pairs are heteromorphic, 0.5 if all pairs are semi-isomorphic and 0 if all pairs are isomorphic.

The classification of the translation pairs in the four semantic categories forms the basis for the following measures of semantic correspondence:

$$\begin{aligned} \text{Semantic Equivalence Index (SE):} \\ EQ/(EQ+MSP+LSP+OTH) \\ \text{Specification Index} \\ (MSP - LSP)/(EQ+MSP+LSP+OTH) \end{aligned}$$

The *SE* measure tells the proportion of sentences with unchanged meaning in the translations while the *Specification index* will highlight the tendency of a translation to be more explicit or more general than the source text.

4. Results

The correspondence model described in section 2 has been applied to the LTC corpus. Each bitext in the corpus was analysed by randomly picking out 100 source sentences with their corresponding target sentences. All non-obligatory shifts were recorded in an SGML-framework, together with the structural and semantic correspondence characteristics for each bitext pair. In a first stage, the actual annotation and analysis were made independently of the two authors. In the second stage the two analyses were reviewed jointly and a final version was agreed upon.

Table 3 below gives the distribution of the structural and semantic correspondences for four of the texts in the LTC. Table 4 summarises the results in terms of the measures described in the previous section for the same texts. A more detailed account of the analysis can be found in Merkel (1999).

Table 3. Structural and semantic correspondence expressed as number of translation pairs in the samples

	ACCESS	CLIENT	GORDI-MER	ATIS
Isomorphic	21	38	33	97
Semi-isomorphic	31	21	24	3
Heteromorphic	48	41	43	0
EQ	36	37	37	90
MSPEC	11	6	27	1
LSPEC	24	34	4	0
OTHER	29	23	32	9

Table 4. Measures for structural closeness, semantic equivalence and target specification.

	ACCESS	CLIENT	GORDIMER	ATIS
Structural Change	0.635	0.515	0.55	0.015
Semantic equiv.	0.36	0.37	0.37	0.90
Specification	-0.13	-0.28	0.23	0.01

The figures clearly demonstrate the differences in terms of change of structure and meaning for these texts. The MT-translated ATIS text is by far the most literal translation both as far as structure and meaning are concerned. 97 of the 100 translations are isomorphic and over 90 are semantically equivalent. The purely human translation of a computer manual (ACCESS) only retains isomorphic correspondence for 21 of the translated sentences and 36 sentence pairs are judged as equal from a semantic point of view. Another interesting difference is that the fiction translation (GORDIMER) exhibits a clear tendency towards semantic specification, whereas the translations of the manuals (ACCESS and CLIENT) both go in the opposite direction, but with different strengths. In Figure 1, the different characteristics of four of the translations in LTC are depicted graphically.

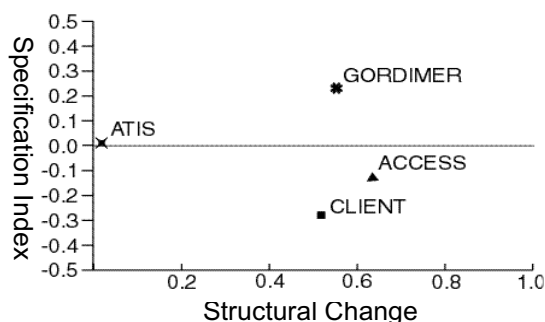


Figure 1. Four different translations displayed according to their tendency for structural and semantic change.

5. Comparison with standard methods of MT evaluation

At least since the beginning of the nineties there has been a growing interest in evaluation of all types of NLP applications, including MT systems, cf. Sparck Jones & Galliers (1995).

Linguistic assessment of MT systems includes measures both for accuracy (or fidelity) and acceptability. Measures for accuracy are oriented to the relation between the translation and the source text, while measures for acceptability focus on the properties of the translation as a text in the target language. Measures of accuracy tend to concentrate on the meaning relations between source and target texts, while measures of acceptability often pays attention both to content and form. A case in point is the DARPA machine translation evaluation paradigm that makes a distinction between the informativeness and the fluency of the translation (White et al., 1994).

The basis for evaluation of properties such as these is often subjective grading. Confidence in the results is obtained by using experts or well-educated subjects in

large numbers who are requested to grade the translations in various respects on a scale.

A simpler and apparently more objective approach is to perform an error analysis of the translation. However, as Sparck Jones & Galliers (1995) point out, the definition of what is to count as an error is not always straightforward and what is an error in one situation may be acceptable in another situation. Also, as is evident from the results in Tables 3 and 4, structural and semantic shifts are parts and parcel of high-quality translations and must not be mistaken for errors. In fact, the translation with the highest number of errors of these four is the closest one, i.e., the translation produced by MT.

The correspondence model, although concerned with the relation between source text and translation, does not provide a measure of accuracy. On the contrary, the aim is to be descriptive and to avoid subjective judgements as far as possible by providing definitions of the different types of shifts and elaborate manuals for annotators. The model is thus complementary to evaluations concerned with accuracy. The value of the model is its ability to identify and quantify different styles of translations, whether accurate or not.

Newmark's (1988) description of source vs. target emphasis as a scale from *Word-for-word translation* to *Adaptation* is a good starting point for characterising translation styles. In between the above endpoints we find translation styles such as *Literal translation*, *Faithful translation*, *Semantic translation*, *Communicative translation*, *Idiomatic translation* and *Free translation*. It seems a reasonable assumption that as we move along this scale, the value of the Structural Change Index would tend to increase.

The issue whether a certain style of translation is appropriate for a given text is one that requires informed subjective judgement. It is possible, however, to measure translations that are considered to be of high quality for a given text type and in this way provide a fine-grained description of the norms that are valid for translations of this type. If there is little variation in the translations the indices may even form the basis for benchmarking. In any case, given such standard norms, any translation, including MT output, may be measured and compared to the norms.

6. Incorporating error analysis

As explained in the previous section the correspondence model does not in itself provide a comprehensive description of the relation between source and target text. To accomplish this the issue of accuracy has to be taken into account one way or another. We have no basis for recommending any particular approach, but when a correspondence analysis is performed, it seems reasonable to pick one that as far as possible uses the same underlying distinctions. For an overview of some previous approaches, see Sager (1993).

Some categories of the correspondence model indicate the presence of an error. This is the case for the semantic correspondence category OTHER and the related lexical category "nucleus with a different meaning". In Table 2, we saw that the MT-translated ATIS text had nine of the translation analysed as OTHER. In these translation pairs two major error types can be distinguished:

1. Incorrect lexical choice (6 instances)
2. Incorrect choice of construction (3 instances)

A case of incorrect lexical choice is the following:

SOURCE: *I would like to **depart** after three p.m.*

TARGET: *Jag skulle vilja **avgå** efter femton.*

Here, the verb *depart* has been translated with *avgå*. This would have been appropriate with an inanimate subject, but when used with a human subject *avgå* means 'resign'. The verb *åka* ('go') would have been a better alternative.

The second error category can be illustrated with the following example:

SOURCE: ***What are the arrival times** in Washington?*

TARGET: ***Vad finns det för ankomsttider** i Washington?*

In many contexts the translation (*what are...->vad finns det för...*) is appropriate, but not in this one. In this case, the combination with "arrival times", would require another rendering to be grammatical in Swedish, e.g. *Vilka är ankomsttiderna till Washington?* (*what are the arrival times in Washington*) eller *När är man framme i Washington* (*when do you arrive in Washington*)?

Other cases of error, such as missing obligatory shifts, e.g., obligatory changes of word order, are not registered in the current model, but obviously such errors must be taken into account in evaluations of accuracy. Consider for example the missing obligatory subject-verb order shift in the following constructed example²:

SOURCE: *On Thursday **I want to go to Chicago.***

TARGET: *På torsdag **jag vill åka till Chicago.***

The above examples are of course very limited as they come from only one short MT translation. Nevertheless, the evaluation indicates the kind of error classification that can complement the current model in order to present a more comprehensive picture of translation adequacy.

7. Summary and conclusion

The major advantages of the correspondence model as a part of MT evaluation are the following:

- It is descriptive and objective. Of course we cannot expect 100% agreement between annotators, but with training and experience it is possible to come quite close.
- It is able to reveal interesting differences between translations that are produced by different methods, e.g., comparing MT systems that use different technologies and, more generally, compare translations produced by MT, MAHT and HT.
- It offers descriptions at different levels of granularity; we may aggregate shifts of the same general type as was done above, but we may also provide measures for individual shifts. In fact, many interesting measures can be derived that we have not been able to discuss for reasons of space limitations.
- It is general; some of the measures that can be generated within the model seem to correlate well

with categories for translation styles that have been developed within the field of descriptive translation studies and translation theory.

At the present time, we do not know how the measures correlate with translation norms and subjective evaluation. This is something that could be tested, however. For example, translations from different relevant genres that satisfy given criteria according to human judgement can be measured on the basis of the correspondence model to find out whether the measures are stable and to what extent they correlate with subjective qualities. In an ongoing Master's project a study is made on the correlation between measures and reading comprehension of translations.

The model as originally applied to the LTC did not register obligatory shifts, the reason being that human translators seldom miss them. If it were to be used in conjunction with evaluation of accuracy, it would be necessary at least to register the obligatory shifts that have been missed in the translation.

An obvious drawback with the method is that it requires quite a lot of human effort. Progress in parsing, alignment and bilingual lexical resources such as EuroWordNet may however enable partial automation of the scoring process.

8. References

- Agnäs, M.-S., H. Alshawi, D. Carter, K. Ceder, M. Collins, R. Crouch, V. Digalakis, B. Ekholm, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, P. Price, S. Pulman, M. Rayner, C. Samuelsson and T. Svensson . *Spoken Language Translator: First-Year Report*. Stockholm, Swedish Institute of Computer Science, 1994.
- Ahrenberg, L. & M. Merkel: Språkliga effekter av översättningssystem. In O. Josephson (ed.) *Svenskan i IT-samhället*. Uppsala, Hallgren & Fallgren: 96-116, 1997.
- Merkel, M. *Understanding and enhancing translation by parallel text processing*. Linköping Studies in Science and Technology, Dissertation No. 607. Department of Computer and Information Science, Linköping University, 1999.
- Newmark, P. *A Textbook of Translation*, Prentice Hall, London, 1988.
- Sager, J.C. *Language Engineering and Translation. Consequences of Automation*. Benjamins, Amsterdam, 1993.
- Sparck Jones, K. & J.R. Galliers. *Evaluating Natural Language Processing Systems. An Analysis and Review*. Springer Verlag, Berlin, 1995.
- van Leuven-Zwart, K.M: Translation and Original: Similarities and Dissimilarities, II. *Target 2(1)*: 69-95, 1990.
- White, J. S., T. O'Connell & F. O'Mara. The ARPA MT Evaluation methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings from the Association of Machine Translation in the Americas (AMTA-94), Columbia, Maryland*: 193-205, 1994.

² This type of translation error is in fact not present in the ATIS translation.