Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory-Systemen

- Ein Erfahrungsbericht -

Uwe Reinke
Institute for Applied Linguistics, Translation and Interpreting
Saarland University
PO Box 15 11 50
D-66041 Saarbrücken
u.reinke@rz.uni-sb.de

Extended Abstract

The paper discusses criteria for evaluating the retrieval performance of translation memories (TM) and describes the results of some corpus-based experiments. Although commercial TM systems have been developed since the late 1980s and TM software is now being widely used in the translation industry, researchers have so far paid rather little attention to this type of tool. Except for the suggestions contained in the EAGLES working group's final report on the 'Evaluation of Natural Language Processing Systems' [Eagles96 – for references see IAI-2000-Bibliog] there are hardly any other proposals for evaluation procedures.

The introductory section of the paper stresses that TMs have much more in common with information retrieval (IR) systems than with machine (MT) translation programms. In contrast with MT systems, TMs do not create any target language (TL) sentences but basically store pairs of source and target language segments-i.e., source language (SL) units (usually sentences) together with their human translations-in order to retrieve them for reuse in identical or 'similar' translation contexts. Thus, in TM systems data processing means data retrieval, while in MT it comprises some kind of data production. This difference also implies a need for different evaluation criteria. Evaluating the performance of a TM involves the measuring of its recall and precision. From the linguistic viewpoint, an evaluation of TMs will deal with intralingual matters (treatment of morpho-syntactic modifications, paraphrases, etc.). When evaluating the performance of MT systems, on the other hand, the focus is on 'translation quality' including such key issues as error analysis, as well as understandability and grammaticality of the TL product.

The second section of the paper takes a brief look at the EAGLES evaluation criteria and discusses some of their shortcomings. The EAGLES group has presented two different benchmarks-one for 'exact matches' and one for 'fuzzy

matches'. The benchmark test for 'exact matches' includes the following steps (cf. [EAGLES96:154]):

- Creation of a text corpus *T* containing texts of the same text type and subject field
- Creation of a TM from a subset of *T*
- Applying the TM to other members of *T*
- Calculating the percentage of translated and correctly translated segments and computing scores based on recall and precision.

The major shortcomings of this test scenario are:

- Selection of text material: In order to reach a sufficient amount of 'exact matches' in a text corpus for TM evaluation, it is not enough to choose texts that belong to identical text types and subject fields. Rather the corpus should be made up of text pairs in which one text is derived from another text in the same language (e.g., 'original' and 'corrected version', 'original' and 'update', 'original' and 'abstract').
- Evaluation of results: As TMs do not translate, it is not the 'correctness of the retrieved translations' that is to be evaluated but the 'relevance' of the retrieved SL/TL segment pairs.

While analyzing the retrieval efficiency for 'exact matches' seems to be a rather trivial task, developing criteria for an evaluation of 'fuzzy match' algorithms is more demanding. The EAGLES group suggests the following benchmarking scenario for 'fuzzy matches' (cf. [EAGLES96:155]):

- Creation of a TM from an authentic text
- Creation of test suites by systematically modifying the material (e.g., changes of punctuation and in numbers and names, changes in segment length, lexical modifications (substitution, addition, deletion of lexical items), changes in sentence structure (word order, grammatical construction))
- Measuring the system's recall.

Although the general outline of this scenario is straightforward, it still offers little help for the evaluation of 'fuzzy match' algorithms because it remains rather vague. While most of the few modifications explicitly mentioned in the report are rather simple and usually do not cause serious retrieval problems (cf. [Rei94], [Rös/War97]), the report does not specify more complex syntactic and semantic variations.

Based on the notion that TMs have much more in common with IR programs than with MT systems, the third section of the paper deals with the question how typical metrics used in IR can be applied to the evaluation of TMs. At first sight, applying 'quantitative measures' like recall and precision to the needs of TM evaluation seems to be a rather straightforward task. Thus, definitions for recall

(*R*) and precision (*P*) can easily be derived from the standard definitions used in IR (cf. [Salt/McG87]):

$$R = \frac{\text{Number of relevant SL / TL segment pairs retrieved}}{\text{Number of all relevant SL / TL segment pairs in the TM}}$$

Yet, a number of questions arise when it comes to defining concepts like 'relevance' and 'similarity'. In IR relevance is seen as the degree of 'formal correspondence' between queries and retrieved documents as well as the degree to which a retrieved document corresponds with the user's information needs [Salt/McG87:173f.]. On the one hand, the 'relevance' of a TM match could be easily measured by comparing the SL query segment and the SL segment retrieved from the TM and counting the number of deletions, additions, replacements, and movements, so that the 'relevance' of a match would decrease with a growing number of 'differences' between query and retrieved segment. On the other hand, translators' 'information needs' could be described as retrieving from a TM an SL/TL segment pair that has the same or at least a 'similar' *content* as the SL segment currently to be translated, so that probably the TL part of the retrieved TM unit can easily be incorporated into the TL text. Thus, defining the 'relevance' of a 'fuzzy match' from the point of view of the translator's 'information needs' requires taking a closer look at the concept of 'similarity'.

For a first distinction of 'similarity classes' it might be helpful to use concepts from 'traditional' linguistics and draw a distinction between 'formal', 'semantic' and 'pragmatic similarity'. The distinction between formal and semantic similarity also occurs in cognitive psychology in connection with experiments on verbal learning and remembering (cf. [Hall71], [USG96]). While for 'formal similarity' distinctive and common features are immediately derived from the 'surface' of the objects to be compared, 'semantic similarity' depends on content-related features. The matching algorithms of today's TM systems basically rely on formal-or more precisely orthographical-similarity. Translators' similarity judgments, however, are mainly based on semantic and pragmatic features.

Following [USG96] the paper divides 'semantic similarity' into 'similarity in meaning' ('Bedeutungsähnlichkeit') and 'conceptual similarity' ('konzeptuelle Ähnlichkeit'). 'Similarity in meaning' implies that the compared linguistic expressions can be substituted without changing the content [1]; 'conceptual similarity' refers to semantic relations like hyperonymy, hyponymy, co-hyponymy, or antonymy. 'Similarity in meaning' comprises both paraphrases and variations in explicitness (i.e., implications and explications). 'Pragmatic similarity' includes such distinctive features like sender-receiver-relation or communicative level. If a SL segment to be translated and the SL part of a TM unit are semantically identical but differ with respect to pragmatic features, adjusting the TL part of the TM unit is unlikely to be worthwhile and in many cases will probably even be impossible.

The fourth section of the paper presents the results of some small-scale corpus-based experiments that try to apply typical IR metrics to TMs. The corpus contains five pairs of German texts that belong to the technical description of a mobile communication system. Each pair is made up of an 'original' and an 'update' text. In the 'originals' subset there are 876 segments (approx. 9,900 words), while the 'updates' subset comprises 898 segments (approx. 11,100 words). To identify the related text fragments of each text pair, all 'updates' were aligned with their 'originals' with the help of a commercial alignment tool. After manually correcting the output of the alignment process and deletion of invariant passages, 126 segment pairs containing syntactically and/or semantically modified fragments were kept as testing material. The segments extracted from the 'originals' subset were used to build TMs for the three TM systems included in the test (i.e., IBM TranslationManager, Star Transit and Trados Translators' Workbench). Then, the segments from the 'updates' subset were used as 'queries'. The total numbers of retrieved segments and retrieved relevant segments were taken to calculate recall, precision, silence, noise and f-measure (i.e., the harmonic mean of recall and precision; cf. [vRij79]). While at first sight, there seem to be great differences between the three systems-with f-measure values ranging from 0.58 to 0.77-these results have to be taken with care because more than 50% of the segment pairs used in the test contain multiple modifications, some of them being so complex that the 'original' is likely to be of little help for translating the 'update' (see the examples in table 4). For this reason, the 66 segment pairs with multiple modifications were used to build a new set of 189 'normalized' segment pairs with each pair containing only one modification. The complexity of the modifications could still vary from e.g. adding a concept in an enumeration to adding a new clause to a composite sentence. In a retrieval test using the 'normalized' segments the results for the three TM systems were much more homogeneous-with f-measures between 0.87 and 0.95. However, a closer analysis of the individual matches reveals that often the systems' match (i.e., 'similarity') values for comparatively 'simple' variations are rather low (see the examples in table 6).

Basically the investigations have shown that

- 'quantitative' metrics like recall and precision alone are too superficial to evaluate the retrieval efficiency of TM systems and require a definition of such 'qualitative' concepts as 'relevance' and 'similarity'
- a 'qualitative' evaluation needs some kind of typology of 'similarity features'
- the use of authentic texts (i.e., pairs of related texts like 'originals' and 'updates') can be difficult in so far as not all of the textual modifications between two related texts are necessarily relevant to evalute the retrieval efficiency of TM systems.

In TM systems retrieval problems mainly occur

- if-in relation to the segment length-there is a comparatively large amount of

different modifications (including variations of the surface structure due to morpho-syntactic modifications)
if segment lengths themselves vary considerably.

While it is likely that retrieval problems caused by morpho-syntactic variations, different word formation patterns or simple syntactic modifications could easily be solved by including common linguistic processes like lemmatisation and determination of derivation patterns, difficulties resulting from differences in segment lengths require more complex procedures for retrieving sub-sentential units (cf. [Rei99]).

Note

[1] Of course, 'similarity in meaning' is a very unfortunate term. First, it is not really suited to denote a sub-category of 'semantic similarity' because it is not a hyponym but rather a synonym of this term. Secondly, if the expressions to be compared are substitutable, then 'meaning'-at least in the sense of 'propositional meaning'-is not a distinctive feature but an invariant. Thus, 'similarity in expression' ('Ausdrucksähnlichkeit') might perhaps be a more appropriate term.