

Linking Translation Memories with Example-Based Machine Translation

Michael Carl and Silvia Hansen

Institut für Angewandte Informationsforschung,
Martin-Luther-Straße 14, 66111 Saarbrücken, Germany,
carl@iai.uni-sb.de

Abstract The paper reports on experiments which compare the translation outcome of three corpus-based MT systems, a string-based translation memory (STM), a lexeme-based translation memory (LTM) and the example-based machine translation (EBMT) system EDGAR. We use a fully automatic evaluation method to compare the outcome of each MT system and discuss the results. We investigate the benefits for the linkage of different MT strategies such as TM-systems and EBMT systems.

1 Introduction

A number of different MT paradigms and systems are described in the research literature and are available on the market. Whereas each system has its strength, none of them leads to an overall satisfactory result when applied to real world texts (cf. (Nübel and Seewald-Heeg, 1998))

The more an MT system generalizes the text to be translated, the more it is likely to achieve a broad coverage. On the other hand, the more it depends on the mere surface form of the translation text, the more it is likely to achieve a high translation precision.

In this paper we give empirical evidence for this hypothesis by comparing the translation outcome of three corpus-based MT systems which use gradually more abstract representations. We use two representational forms in a translation memory, a string-based translation memory (STM) and a lexeme-based translation memory (LTM) and the example-based machine translation system EDGAR (for an in depth description of EDGAR see (Carl, 1999) in these proceedings).

Both TMs make use of the FindLink database retrieval software distributed by CONNEX (cf. (CON, 1996; Heitland, 1994)). FindLink is also used by commercial TM manufacturers such as TRADOS (Translator's Workbench), STAR (TRANSIT) and ZERES (Zer, 1997). In the learning phase, FindLink stores a set of reference translations in its database while coding the match string (i.e. the source side of the translation) into an n-gram sequence. In the translation

phase, the search string (i.e. the translation sentence) is coded in the same way and the translation(s) of the most similar match string(s) are returned as best (available) translations. Each returned translation is associated a match score M between 0% and 100% which indicates the similarity of the search string and the match string.

In the STM the surface forms of the reference translation's source language sides are used as a match string, whereas the LTM match strings are based on the lexemes of the reference translation's source language content words. The representation in the LTM is thus an abstract of inflection and derivation, while the STM stores the surface forms of the match strings without performing any abstraction. Some TMs such as Translator's Workbench and TRANSIT follow the STM approach. The ZERES implementation is a mixture of the STM and the LTM approaches.

EBMT system EDGAR relies on morphologic analysis of both languages involved and on the induction of translation templates from the analyzed reference translations. EDGAR decomposes the translation text at several levels of generalization by matching it against translation examples contained in a case base. The matched chunks are then specified and refined in the target language. Among the three MT systems under consideration, EDGAR uses the most generalized representations and is thus expected to have the broadest coverage, while the STM uses the least generalized representations and is therefore expected to yield most precise translations.

In order to test this hypothesis empirically, we have used two bilingually aligned translation corpora: a reference corpus and a test corpus. Each system was trained with a reference corpus containing 303 German-English translation examples as produced by a car manufacturer. The test corpus contains 265 translation examples from the same company and the same sub-domain (repair instructions). The size of the sentences ranges from 1 up to 160 characters in length containing single numbers and word translations, short imperative sentences, noun phrases and

whole sentences with subordinate clauses. The test corpus and the reference corpus are from the same domain, with similar vocabulary and similar phrase structure.

To train EBMT system EDGAR, the reference corpus was first sub-sententially aligned. The generated set of translation examples was generalized and compiled into an EDGAR case base which was used for translation.

For each of the systems we carried out two translation tests. To verify the reliability of each system, we first translated the reference text and compared the output with the ideal translation contained in the reference corpus. Then, the test text was translated and compared with the ideal translation contained in the test corpus.

From the reported experiments we conclude that a linkage of different MT paradigms such as TM technologies and EBMT systems may enhance the overall translation result.

2 Translation Scores T^1 and T^2

A great variety of methods have been proposed in the literature to quantify the quality of MT systems but yet there is no general agreement on MT evaluation methodology. This is partly due to the problem of the “ideal” translation (i.e. to decide which of a number of possible translations is the “best” one) and partly due to the state-of-the-art in MT (i.e. one cannot expect a high quality all purpose MT system). Evaluation methods in the last few years examined the structure and complexity of the input and generated output text, recent research in MT evaluation has shifted to a “task-diagnostic approach” (Vanni, 1998) which focuses on the applicability of the generated translation text.

Roughly, one can distinguish between manual evaluation methods and (fully) automatic evaluation methods. In the former case a bilingually skilled person checks the output of the translation system and validates the translation outcome according to pre-defined set of quality criteria (cf. e.g. (Nübel and Seewald-Heeg, 1998)).

In a fully automatic evaluation scenario, translation quality criteria are fully formalized such that a computer program can check the translation result without the need for a human supervisor. Such method usually makes use of a test corpus to check the generated translations against an “ideal” translation contained in the test corpus. Again, several measures have been proposed to grade the translation quality automatically. These measures compare the generated translation string with the ideal translation string and quantify the difference between the two.

In (Meyers et al., 1998) it is claimed that fully automatic evaluation methods can be used to validate

enhancement efforts in MT systems, and it reensure that incremental changes of a system are for the better rather than for the worse. Meyers et al. (1998) proposes an evaluation procedure which computes the ratio between the complement of the intersection set of the generated translation T and the ideal translation I ($|\overline{I \cap T}|$) and the combined length of these two sentences ($|I + T|$) as shown in equation 1.

$$TS^2 = \frac{|\overline{I \cap T}|}{|I + T|} \quad (1)$$

A generated translation T which has no word in common with the ideal translation I has a translation score 1 because the complement of the intersection $|\overline{I \cap T}|$ is equivalent to the length of the concatenation $|I+T|$. On the other hand, if the generated translation is identical to the ideal translation the complement of their intersection is empty and thus, the translation score yields 0. We refer to this translation score as the TS^2 .

In addition to the TS^2 translation score we use a slightly different TS^1 translation score as follows.

$$TS^1 = \frac{|T \cap I| * \min(I, T)}{|T| * \max(I, T)} * 100\% \quad (2)$$

The expression $|T \cap I|$ denotes the size of the intersection of the generated translation T and the ideal translation I . $|T|$ is the number of lexemes in the generated translation T and $\min(I, T)$ and $\max(I, T)$ are the number of lexemes of the shorter and longer sentences I and T respectively. The measure yields 100% if the generated translation T has the same content words and both have the same length. It yields 0% if the generated translation and the ideal translation have no content word lexemes in common.

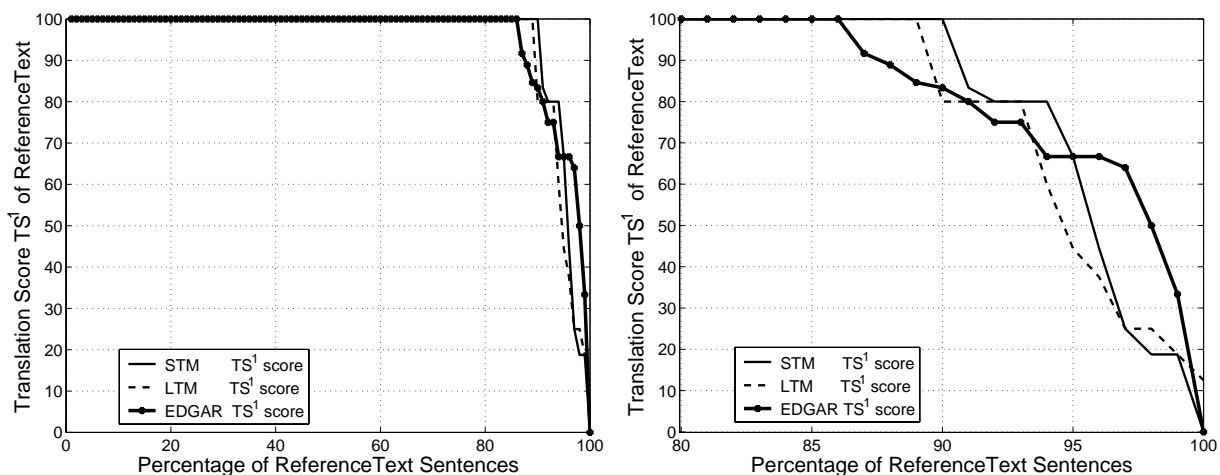
This measure penalizes translations when the generated translation is longer than the ideal translation slightly more. For instance, a generated translation $T : abcdf$ mapped against the ideal translation $I : abc f$ achieves a translation score of 64.0% whereas the generated translation $T : abc f$ and an ideal translation string $I : abcdf$ achieves a translation score of 80.0%. The translation score TS^1 has the advantage that it can directly compare with the match score M which indicates the self-estimation of the TMs translation success rate.

For the calculation of both translation scores TS^1 and TS^2 only lexemes of the content words are considered. Although the shapes of the curves are slightly different, both translation scores have the same implications.

3 Translation of Reference Text

To verify the reliability of each system, we have translated the reference text (i.e. the text corpus used

Figure 1: Translation score of Reference Corpus



The curves put the generated reference corpus translations in relation to their achieved translation scores TS^2 for the three MT systems. The right picture is an extract of the left picture showing the translation scores between 80% and 100% the translated sentences.

to train the system) and compared the output with the reference translations. The diagram in Figure 1 illustrates the translation results for the three systems.

None of the systems translates the reference text entirely correctly. Best results are obtained with the STM where 275 (or 90.7%) of the 303 reference achieved a translation score of 100%. Worst performed EDGAR which only translated 262 (85.5%) of the reference sentences a 100% correctly. The lexeme-based TM achieved for 89.4% (271 sentences) a translation score of 100%. EDGAR performs best with decreasing translation score. 96.6% or 293 sentences of the EDGAR translations achieved a translation score of 66% or better while the lexeme and the STM translates 93.7% (284 sentences) and 95.0% (288 sentences) respectively with the same translation score.

Initially, one would expect that all systems generate only correct translations for the text they have been trained on. Malperformance is due to a couple of reasons. Ambiguous translation examples contained in the reference text are an important factor. Ambiguous translations of one source language expression do not guarantee that the system chooses the ideal one in the translation process. For instance, the reference corpus contained three English translations for the German sentence “ENTER betätigen.” as shown in the table 3. If we take the ideal translation to be “Press ENTER on the PDU screen.” the generated translation “Select ENTER.” would yield a translation score TS^2 of 25.0%. If we take the ideal translation to be “Press ENTER on the display” the translation score for the same generated translation would yield 33.33%. While in both cases the generated translation has two content words, the former ideal transla-

tion contains three content words and the latter ideal translation has four content words. Each of the ideal translations shares one content word with the generated translations. A few such ‘lexical’ ambiguities found in the reference text are shown in table 3.

A second reason which introduces ‘ambiguity’ in the TM-based translations is the underlying retrieval software¹ used for the experiments. The fuzzy matching algorithm of the retrieval software returns 100% match score M if the search string is a substring of the match string contained in the database. In this way, the German search string “Schalter D nach 4” returns a match score of 100% for all entries listed in table 4 below. We have, however, filtered the retrieved output in such a way that only match strings are returned which are at most twice as long as the search string. Note also that the word order does not affect the retrieval result. Thus, even though “4” and “D” are permuted in the fifth match string in table 4, the retrieval software returns 100% match.

In some cases function words can help to distinguish between different but similar search strings as shown in table 6 below. The function word “Der” in the search string “Der Kickdown-Schalter” makes the retrieval of the ‘best’ translation possible.

In the lexeme-based TM it is only the lexemes of content words that are stored while function words are discarded from the match and from the search strings. Lexemization of content words transforms the search string “Der Kickdown-Schalter” into “Kickdown Schalten” so that all of the match strings obtain

¹As previously noted, we use the FindLink software (cf. (CON, 1996; Heitland, 1994)). Commercial TM manufacturers using this software such as TRADOS, STAR and ZERES have implemented further filters to avoid some of the ambiguities.

Lexical translation ambiguities found in the reference text.

German sentence		English Translation	(3)
Enter betätigen.	↔	{ Select ENTER Press ENTER on the PDU screen. Press ENTER on the display	
Gestängehebel	↔	{ Gear shift lever. Transmission Unit Gear Selector	
Massekabel von Batterie abklemmen.	↔	{ Disconnect the battery ground lead Disconnect the vehicle battery ground lead.	
Aussenseil befestigen.	↔	{ Locate the outer cable Secure the outer cable	

Ambiguities 'generated' by the match score M of the retrieval software.

search string	M score	match string	(4)
Schalter D nach 4	100% →	{ Steckverbinder Schalter D nach 4 Der Schalter D nach 4 Einbaulage Schalter D nach 4 Schalter D nach 4 Schalter 4 nach D Schalter D nach 4 ausbauen.	

Ambiguities 'generated' by the LTM when matching the lexemized forms of the content words only.

search string	M score	match string	(5)
Kickdown Schalten	100.00% →	{ Kickdown Schalten vorhanden Kickdown Schalten Kickdown Schalten Kickdown Schalten ausbauen Kickdown Schalten prüfen einstellen	

a 100% match score as shown in table 5. The table 5 shows lexeme representations of the match strings from table 6.

In this way lexemes introduce further ambiguities which are not resolved in the translation process. Since only the translation of the first (of possibly many) best match strings is considered for the translation scoring, the lexeme representation cannot compete with the string representation when translating the reference text. Whereas the STM retrieves the exact translation and thus achieves 100.00% translation score, the LTM only obtains 44.44% translation score for the reference translation "*Der Kickdown-Schalter* ↔ *The kickdown switch*". This is shown in the table 8.

A different kind of ambiguity comes into play when looking at the EDGAR translations. EDGAR decomposes and generalizes the sentence to be translated and specifies and refines the generalization in the target language. Decomposition is based on reference translation examples and sub-sententially aligned translation segments extracted thereof. This produces further translation ambiguities because the system must chose among (probably inconsistent) translation equiv-

alences generated in the alignment step.

Table 7 shows a three translation examples extracted from the reference corpus in the alignment step. Applying the first translation example *Gangwahl und Schaltposition* ↔ *Selection & Gear Position* yields a translation score of 75.00% for the EDGAR translation whereas both of the LTM and the STM achieve 100.00% (depicted in table 9).

Another problem occurs in morphological synthesis when generating the target language surface string from the lexeme representation. Because one (of two) lexeme representation for the word "two" is "2", the morphological synthesis program was unable to generate the appropriate surface string (i.e. *two*).

4 Translation of Test Text

The diagrams in Figure 2 (left) depict the translation scores of the test text. The vertical axis represents the translation scores whereas the horizontal one shows (in percent) the 265 translations of the test text. The upper graphics depict the TS^1 translation scores,

The lexemization ambiguities in 5 do not appear in the STM when matching function words.

search string	M score	match string	(6)
Der Kickdown-Schalter	86.67% →	Kickdown-Schalter (wenn vorhanden) und	
	100.00% →	Der Kickdown-Schalter	
	83.53% →	Kickdown-Schalter	
	83.53% →	Kickdown-Schalter ausbauen.	
	83.53% →	Kickdown-Schalter prüfen und einstellen	

Ambiguities introduced in EDGAR's subsentential alignment process.

German aligned side		English aligned side	(7)
Gangwahl und Schaltposition	↔	Selection & Gear Position	
Gangwahl und	↔	Shift Selection &	
Gangwahl	↔	Selection	

Translation scores TS^1 for the translation:

<i>Der Kickdown-Schalter</i>	→	<i>The kickdown switch:</i>	(8)
EDGAR:	50.00%	The kickdown	
LTM:	44.44%	The kickdown switch (where fitted)	
STM:	100.00%	The kickdown switch:	
<i>Gangwahl und Schaltposition</i>	→	<i>Shift Selection & Gear Position</i>	(9)
EDGAR:	75.00%	Selection & gear position	
LTM:	100.00%	Shift Selection & Gear Position	
STM:	100.00%	Shift Selection & Gear Position	

as for the outcome of the TS^2 translation scores in the lower one. In addition to the translation scores, the upper left graphic plots the match scores M of both TMs. The main properties of the upper and the lower curves are similar, with the difference that the TS^1 score yields slightly steeper degradation of translation performance. As the match scores represent the TMs' self-estimation, the upper graphic is somewhat more TM-friendly because the match scores come closer to the actual translation scores.

The right diagrams in Figure 2 show translation scores (upper TS^1 ; lower TS^2) for the 128 test corpus translations of the STM and EDGAR with an STM match score of 80% or below. The upper graphic includes the STM match scores (beginning at 80%). Again, the upper and the lower curves are quite similar.

In the following we discuss in greater detail the upper graphics plotting the TS^1 scores in more detail.

The upper left diagram in Figure 2 shows that both, STM and LTM, generate more ideal translations than EDGAR. From the 265 sentences of the test corpus, the STM generates 104, the LTM generates 100 and EDGAR generates 96 ideal translations with reference to the translation score TS^1 . This represents

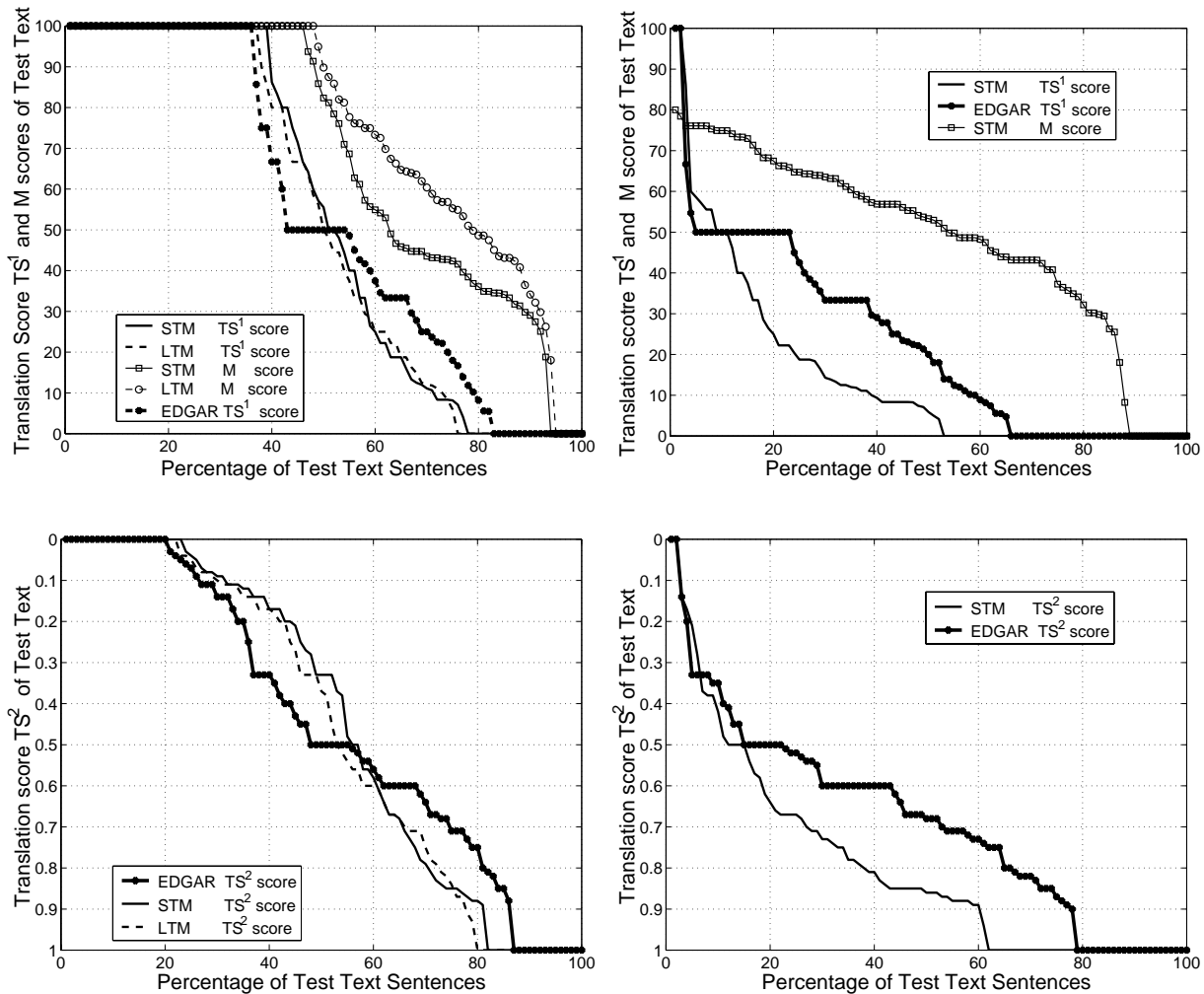
39.2%, 37.7% and 36.2% of the test corpus sentences respectively. Not only are more ideal translations produced by the STM, but the STM self estimation is also more reliable. It assigns to 122 (47.7%) of the sentences a 100% match score whereas the LTM assigns to 129 (or 48.7%) sentences a 100% match score. From the 122 sentences assigned an STM match score of 100%, 102 (i.e. 83.6%) translations achieve a translation score TS^1 of 80% or more, whereas from the 129 sentences assigned the LTM 103 (i.e. 80.0%) a TS^1 score of 80% or more.

For a match score of 80%, which is likely to be chosen by users of translation memories, these numbers are as follows.

	$M > 80\%$	$TS^1 > 80\%$	$TS^1 = 100\%$
STM	137 (51.7%)	109 (79.6%)	94 (68.6%)
LTM	143 (54.0%)	110 (77.0%)	95 (66.4%)

If it comes to less ideal translations, i.e. translations of sentences which are not (or only partially) covered by the reference corpus and where the match score decreases consequently, EDGAR tends to generate better translations than both of the string and lexeme-based TMs. EDGAR generates 144 translations (i.e. 54.5% of the test sentences) achieving a translation score higher than or equal to 50%. LTM

Figure 2: Translation score of Test Corpus



The left diagrams depicts (in percent) translation scores of the 265 test text sentences. The right diagram shows the 128 (STM and EDGAR) translation which achieved 80% STM match score or less. The upper two diagrams shows evaluation of test text translations according to the TS^1 ; the lower according to the TS^2 . In addition to this the upper left diagram plots the LTM and STM match scores and the upper right diagram plots the STM match scores $\leq 80\%$.

translates 134 of the sentences (50.8%) and STM 139 sentences (52.7%) with the same translation score.

While both STM and LTM perform better if near matches can be found in the TM's database, EDGAR achieves better translation scores where no exact match is available in the reference corpus. This can clearly be seen in the right-hand side graphics in Figure 2. These graphics plot the 128 test corpus sentences which achieve a STM match score of 80% or less. EDGAR almost certainly produces better translation scores with respect to both the TS^1 and the TS^2 .

Due to EDGAR's decomposition, generalization and refinement capacities, new translations can be composed from smaller parts. Decomposed chunks are translated where a translation is available in the case base and subsequently refined into the target lan-

guage. Chunks which do not fit an entry in the case base appear as source language strings in the translation. EDGAR may then generate hybrid German-English translations which are only partly translated. Table 10 shows the generated translations for the test sentence "Park position switch \leftrightarrow Parkstellungsschalter" for each translation system.

German compound nouns such as "Parkstellungsschalter" are decomposed into their individual morphemes *park/stellung/schalter* and in absence of a full translation entry are compositionally translated. Translating the test sentence "Stecker - Parkstellungsschalter \leftrightarrow Connector - Park position switch" EDGAR generates the German-English hybrid "Connector - parks stellung schalter" as shown in table 11. It thus achieves 50% translation score whereas both TM

Ideal Translation	:	Park position switch	(10)
EDGAR:	33.33%	parks stellung schalter	
STM:	18.75%	Kickdown Switch - Check & Adjust	
LTM:	18.75%	Kickdown Switch - Check & Adjust	
Ideal Translation	:	Connector - Park position switch	(11)
EDGAR:	50.00%	Connector - parks stellung schalter	
LTM:	16.33%	Ensure that contact with the switch does not result in switch movement	
STM:	16.00%	From above, disconnect the rotary switch harness multiplug	
Ideal Translation	:	Incorrect cable adjustment	(12)
EDGAR:	33.33%	Selector falsch adjusted	
LTM:	66.67%	Selector Cable - Adjust	
STM:	8.33%	Detach the selector cable from the gear shift assembly.	
Ideal Translation	:	$\left\{ \begin{array}{l} \text{Movement of the lever across the gate to 4, 3 and 2 positions} \\ \text{disengages the cable from the selector lever and engages} \\ \text{the DLS which controls gear selection electronically.} \end{array} \right.$	(13)
EDGAR:	42.50%		
LTM:	29.41%	The position of the gear selector lever is detected by the range sensor; a system which consists of two (2) sensors (switch systems).	
STM:	5.88%	Selects Normal or Sport mode when pressed by the driver.	

translations achieve even worse results due to the lack of an appropriate reference example and decomposition possibilities. In this way EDGAR achieves higher translation scores for only partially covered sentences or partially translated compound nouns.

Another example is shown in table 12 where EDGAR generates a hybrid translation for the test sentence “*Seilzug falsch eingestellt* \longleftrightarrow *Incorrect cable adjustment*”. The best translation score here is achieved by LTM.

This situation is still more crucial if the length of the sentences to be translated increases. Since it is less probable for the TM to find good fitting translation examples for long sentences, EDGAR outperforms the other approaches as shown in table 13 when the following sentence is to be translated:

Durch Bewegen des Hebels in die Stellungen 4, 3 und 2 wird der Seilzug vom Wählhebel getrennt und die elektronische Gangwahl durch den Bereichsschalter aktiviert.

\longleftrightarrow

Movement of the lever across the gate to 4, 3 and 2 positions disengages the cable from the selector lever and engages the DLS which controls gear selection electronically.

In table 13, both TM systems yield perfect English sentences. However, their content is completely misleading and it is unlikely that a translator can make any use of such translation proposals. EDGAR achieves the highest translation score and parts of the source sentence may even be properly translated. It is, however, disputable whether the generated translation helps to understand the source language sentence better.

5 Conclusion

This paper compares three gradually more generalizing, corpus based MT systems: the example based MT system EDGAR, a string-based translation memory (STM) and a lexeme-based translation memory (LTM). Our results show that the least generalizing system (the STM) achieved higher translation precision when near matches can be found in the data base. However, if the reference corpus does not contain any similar translation example, EDGAR performed better than the STM and the LTM. We therefore conclude that the more an MT system is able to decompose and generalize the translation sentences, translate parts or

single words of it and to recompose it into a target language sentence, the broader is its coverage and the more it loses translation precision. A combination of different MT approaches thus seems appropriate to improve the overall translation result.

Further experiments shall clarify whether such complementary capacities of MT systems provide a useful help to enhance translation memory performance. For this end we plan to integrate dynamically a commercial STM (i.e. TRANSIT) with the EBMT system EDGAR in such a way that the TM can request translation proposals from the EBMT system in case it falls below a certain threshold.

In order to improve the EBMT system EDGAR and to evaluate its capacities and limits, further experiments involving test texts which are gradually different from the reference corpus are to be conducted in the future. In this way, we try to refine the sub-sentential alignment tool so that ambiguous translations can be excluded and better translation results can be achieved.

References

- Michael Carl. 1999. Inducing Translation Templates for Example-Based Machine Translation. In *MT-Summit VII*.
- CONNEX, Hildesheim, Germany, 1996. *CONNEX: WORKSHOP*.
- Michael Heitland. 1994. *Einsatz der SpaCAM-Technik für ausgewählte Grundaufgaben der Informatik*. Ph.D. thesis, Universität Hildesheim, Fachbereich IV, Hildesheim, Germany.
- Adam Meyers, Roman Yangarber, Ralph Grishman, Catherine Macleod, and Antonio Moreno-Sandoval. 1998. Deriving transfer rules from dominance-preserving alignments. In *Computerm, First Workshop on Computational Terminology*, Montreal, Canada.
- Rita Nübel and Uta Seewald-Heeg, editors. 1998. *Evaluation of the Linguistic Performance of Machine Translation Systems*. Computerlinguistik und neue Medien. Gardez! Verlag, St. Augustin, Bonn.
- Michelle Vanni. 1998. Evaluating MT Systems: Testing and Researching the Feasibility of a Task-Diagnostic Approach. In *Translating and the Computer 20, Aslib*, London, November.
- Zeres GmbH, Bochum, Germany, 1997. *ZERESTRANS Benutzerhandbuch*.