Towards the Integration of MT into a LSP Translation Workflow

David Vilar[1], **Michael Schneider**[2], **Aljoscha Burchardt**[1], **and Thomas Wedde**[2]

[1]DFKI , Language Technology Lab , Berlin, Germany {name.surname}@dfki.de
[2]beo GmbH, Stuttgart, Germany {name.surname}@beo-doc.de

Abstract

This user study reports on an ongoing pilot that aims at using machine translation on a large scale, for the translation of technical documentation for a globally acting automotive supplier. The pilot is conducted by a language service provider and a research institution. First results go beyond expectations.

1 Introduction

In real-world translation environments efficiency, both in terms of cost and time, is of critical importance. Even more when the volume of texts to translate is large. Machine translation (MT) seems to be a good candidate for achieving these goals, but somehow surprisingly the economic feasibility of MT and the fitness for real-world needs of professional translators and Language Service Providers (LSPs) have been hardly analysed so far.

The MT community tries to broaden the domains the translation systems are applied to. In the early years, research on statistical machine translation concentrated on restricted domains, the touristic domain being a typical example. As the quality of the translations got better, the difficulty of the task was increased by moving to richer domains. The WMT evaluations are another example of this trend. In the first editions (Koehn and Monz, 2005) the data the systems were trained and evaluated on consisted only of the proceedings of the European Parliament. In more recent editions (Callison-Burch et al., 2011) the (parallel) training data still mostly consists of europarl data, but the evaluation has moved to the news domain, with a much wider variety of topics.

The goal of this research direction is clear: to produce an "universal translator" that is able to translate any type of text. This is however a very optimistic goal and current systems are still very far from it. And it also may not be the optimal goal for professional translators. When a LSP has a translation request, it is usually accompanied by guidelines of style, vocabulary, etc. Also, the domain is usually quite restricted. Not much topic variation can be expected from, say, user manuals of heavy machinery.

As such, the research community has perhaps overlooked a potential niche where machine translation, in its current development state, can prove to be beneficial. At the same time, potential customers are reluctant when it comes to financing the development of specialised MT engines as their idea is that MT comes for free. In this paper we present a pilot study where we analyze how a state-of-the-art machine translation system performs in a real-life environment. The work is a collaboration between a LSP (beo) providing the experience on real-life translation tasks for Bosch, and a research institution (DFKI) providing the know-how about statistical machine translation.

The paper is structured as follows: Section 2 presents the viewpoint of the LSP on the translation task, as well as the expectation of a machine translation system to be considered useful in their workflow. Section 3 describes the machine translation system adapted for the task. Conclusions are drawn in Section 4.

2 From TM to MT: The LSP's starting point

A LSP always has to keep a good balance between prices, linguistic quality, and time, all for the benefit of the client. Especially in the area of train-

ing material the price pressure is even higher than normal, because professional translation of slides used (internally) for training is routinely omitted. Often the material is created newly in foreign languages if needed, leading to significant differences in content and quality.

The Bosch Automotive Aftermarket department (AA) decided relatively early to have training material translated to keep at least the content consistent among language versions. Price was (and is) still important: translation costs are traditionally not shared with the trainees, and were in fact in the past not part of the training budget.

The net effect was that most of the material did not get translated at all, and if so, without consistently controlled quality.

This was the point when Boschasked beo to take over these translation tasks. Being a "preferred supplier for the Boschgroup" for translation services, it was expected that beo

- keeps the price per word low, at a level of about 70% of the normal word price
- reduces the turnaround time for translations, from 3-4 Months down to ca. 2-4 weeks per unit
- raises the overall quality.

Of course, Translation Memory Technology (TM) was to be used, which helps to keep translations consistent over time and to control terminology and overall quality. But the price pressure is still on: More and more clients are not willing to pay for translation proposals coming from perfect (100%) TM matches. Still, these "synthetic" translations need to be proof read and quality checked by the translator and thus require (paid) work.

It was quite quickly clear to us that a TM alone would not be sufficient to reach the goals, especially the cost limits. When communicating such troubles to clients a common reflex is "why don't you use machine translation?", with the implicit assumption that MT is essentially available for free[1] and with sufficient quality to be used unchecked.

Not so with Bosch. It was known that automatic translations had to go through some sort of quality control, and that MT itself is not free of cost (license costs, machine time, etc.). In this context we came to an agreement to use these training materials as a test case for a pilot project to integrate MT into a professional translation workflow.

The core requirements for this workflow are:

- integration of MT into a traditional TM environment. The translator should be able to use the tools and environment he is accustomed to, to keep productivity high
- no "post editing" of MT results at a large scale. Post editing poses new resource problems as there are usually not enough "post editors" at hand, and they will probably not work for free... Therefore the precedence is translation memory over machine translation over translate from scratch.
- "break even" point for translation costs reachable after roughly 10 months

beo's previous experience with machine translation is limited to post-editing jobs. High volume post editing jobs for different clients lead to the insight that post editing performed as an extra work step is neither cost effective nor a guarantee for good quality. Thus, the objective of the project is to integrate MT in such a way that automatically translated content is "magically" presented to the translator just like a TM match. The translator then is responsible to accept, change or reject the translation, just like a TM match. Standard quality assurance work steps and tools can be applied, the MT is seamlessly integrated into the standard translation workflow along the TM.

3 Training an MT engine

In order to train a translation model, DFKI first had to prepare the data into a format suitable for the translation system. The original format is composed of slides translated from an original language (German) into a target language (in our experiments English and Spanish). The slides themselves could be considered as the translation unit, but we chose to work with sentence-like units. For this we firstly applied an automatic sentence splitting tool, and then proceeded to re-align the produced sentences with the Microsoft bilingual sentence aligner (Moore, 2002).

After some cleanup of the data, including removal of duplicate sentences, a special categorization step has been applied to detect tokens that can

[1] In many (all?) cases "machine translation" is the same as "Google Translate" in the view of the clients.

	DE-EN		DE-ES	
Set	Segm.	Words	Segm.	Words
Original	203K	7.5M	199K	7.3M
Train	402K	3.3M	400K	3M
Dev	2 086	17 746	993	7 790
Test	2 057	16 774	1 008	8 597

Table 1: Statistics of the random split into training, development and test sets. The number of segments in the original data corresponds to slides, in the train, dev and test sets, to sentence-like units.

be directly carried over from the source language to the target language. These categories include numerical quantities, in-text references ("see Table x", legend of Figures, etc.) which are specially marked in the text as well as some formatting information (most notably tabular alignments).

A random split into training, development and test data was carried out. Table 1 shows the statistics of the resulting sets. As can already be seen from these statistics, the data is highly redundant. The number of segments is greatly increased when comparing the original data with the preprocessed data (train, dev and test sets), due to the sentence splitting. On the other hand the number of words is less than half, due to the removal of duplicates.

On this data a phrase-based statistical machine translation system was trained (Zens et al., 2002). We chose the Jane translation toolkit (Vilar et al., 2010) over the more widely known Moses toolkit (Koehn et al., 2007) due to its ability of handling the categories described above.[2] The results in terms of BLEU score are given in Table 2. As can be seen, the scores are very high, around 64%. To give a comparison, the highest scoring system in the 2011 WMT Evaluation Task scored 25% BLEU on the German-English task. For English-Spanish (there was no German-Spanish task) the best scoring system achieves a BLEU score of 35% (Callison-Burch et al., 2011).

The reason of our exceptionally good results lies of course in the nature of the data. As was pointed out before, by its nature the data is highly repetitive, even with sentence duplications removed.[3]

[2] A short note about licensing: Jane is freely available for non-commercial use. At the current stage this study is still of scientific nature. Should a commercial application arise, the licensing issue will have to be reconsidered.

[3] Without removal of duplicated sentences the scores go over 70% BLEU.

Language Pair	BLEU[%]
German-English	64.2
German-Spanish	63.9

Table 2: Results in terms of BLEU score on the test set.

Figure 1 shows some example translations. The first one shows an example sentence where the translation system achieved a perfect translation. The structure of this sentence allows for easy generalization (think of several connector colors) and also shows the categorization carried out when preprocessing the data, where the system detected a number and a reference.

The performance of the system is also quite good for more complicated sentences, as the second example of Figure 1 shows. Although it may sound a bit artificial at first sight due to the repetition of "side" towards the end of the sentence,the automatic translation is actually more accurate than the reference translation and in a technical domain like the one we are dealing with it may be fully acceptable.

Of course not all the translations are good, as the third example shows. Although to be fair to the translation system, this sentence does not fully conform to Bosch's style guidelines (the passive voice should be avoided).

4 Outlook & Conclusions

We have presented a user study of applicability of (statistical) machine translation to a real-life translation task as requested from a LSP. The quality of the resulting translations is very high, well beyond our initial expectations. We consider that the quality is good enough to step to the next phase of the project, integrating the translation system into the human translator's workflow. The goal will be to complement the currently used translation memories, which have proven to be of great assistance to the translator's work. A straightforward application will be to use the translation system when the match of the translation memory is not good enough, but more complex interactions will be considered in a further study.

In the current study machine translation's flexibility to translate phrases like "see Figure 5" even if the number "5" did not occur in the training data has already proven helpful as compared to standard

Source	- Anschlussstecker schwarz (Kl . $number { 31 }) an Buchse $ref { <1> }
Translation	- Black connector (term . 31) to socket <1>
Reference	- Black connector (term . 31) to socket <1>
Source	Werden Sollwerte erreicht , liegt ein Defekt im Airbag-Steuergerät oder im Seitenaufprall-Sensor Beifahrerseite vor .
Translation	If set values are attained , there is a fault in the airbag control unit or in the passenger 's side side impact sensor .
Reference	Airbag control unit or front passenger 's side impact sensor is defective if set values are attained .
Source	Konstruktionsbedingt können auch bei abgebautem Steuergerät keine Wicklungswiderstände gemessen werden .
Translation	The design may also be detached control unit is not winding resistances be measured .
Reference	The design is such that it is not possible to measure winding resistances even with the control unit detached .

Figure 1: Translation examples.

translation memories that present a fuzzy match in these cases. One example of a more complex interaction would be to use machine translation systems for ranking multiple 100%-matches of a translation memory according to plausibility, possibly taking context into account. Once confidence estimations of machine translation systems will get more reliable, human post-editors can be presented only material that needs to be touched or error checked.

Although BLEU scores and inspection of the translations may give a good overview of the translation quality, the final performance test will be of course to measure human performance when using the developed system. The final goal is to improve the efficiency of the whole translation pipeline.

This study may also serve as a hint for the machine translation community. The goal of creating machine translation systems that are capable of dealing with a very wide domain is certainly appealing, but ignoring smaller domains may miss important applications. Our results may seem non-conclusive to some researchers ("too similar training and test data"), but we are dealing with *real-life* data, provided by a LSP. The fact that translation memories are the most widely used computer aid by human translators is an indication that such conditions are realistic.

References

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Koehn, Philipp and Christof Monz. 2005. Shared task: Statistical machine translation between European languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic, June.

Moore, Robert. 2002. Fast and accurate sentence alignment of bilingual corpora. In Richardson, Stephen, editor, *Machine Translation: From Research to Real Users*, volume 2499 of *Lecture Notes in Computer Science*, pages 135–144. Springer Berlin / Heidelberg. 10.1007/3-540-45820-4_14.

Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon Models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 262–270, Uppsala, Sweden, July. Association for Computational Linguistics.

Zens, Richard, Franz Josef Och, and Hermann Ney. 2002. Phrase-Based Statistical Machine Translation. In *German Conference on Artificial Intelligence*, pages 18–32, Aachen, Germany, September.