Using Statistical Machine Translation for Computer-Aided Translation at the European Commission

Andreas Eisele
Directorate-General for Translation
of the European Commission
Luxembourg
andreas.eisele@ec.europa.eu

Caroline Lavecchia
Centre de Recherche Public
Henri Tudor
Luxembourg
caroline.lavecchia@tudor.lu

Abstract

This paper[1] describes recent developments and work in progress in the project MT@EC at the Directorate-General for Translation (DGT) of the European Commission, towards providing large-scale machine translation (MT) functionality for institutions of the European Union.

We present the overall scope of this project and some of the steps taken so far, including an implementation of prototypes based on statistical MT (SMT), we describe various attempts to assess the quality of these prototypes, as well as a first integration of these prototypes into the translation workflow at our institution.

We then sketch ongoing work towards an integration of SMT with translation memory lookup along the lines given in (Koehn and Senellart, 2010) and present first results based on real-life data.

1 Overview

With 27 Member States and so far 23 official languages, the institutions of the European Union face an ever increasing demand in translation of legal, administrative and political documents. According to Regulation No 1 from 1958 and its various amendments, documents of the EU may be drafted and published in any of the official languages. Furthermore, citizens may communicate with EU institutions in any of the official languages, and the institutions are obliged to respond in the same language.

[1]Responsibility for the information and views set out in this report lies entirely with the authors.

The Commission introduced MT in the 1970s as a means to increase translation capacity and to obtain quick translations for comprehension purposes. Over this period a rule-based system was developed that ended up covering 10 EU languages and 28 pairs in December 2010, when it was suspended. In the last few years the publication of the EuroParl corpus (Koehn, 2005), regular annual workshops on SMT with open evaluation campaigns (Koehn et al., 2005)...(Callison-Burch et al., 2011), the emergence of the open-source SMT toolkit Moses (Koehn et al., 2007) and projects like EuroMatrix and EuroMatrixPlus providing Community funding for such activities have started to create an infrastructure focused on making statistical and hybrid MT technologies effective for the translation into a broad set of EU languages.

In this context the Commission started working since June 2010 on building a new MT service, called MT@EC, which should be flexible both in terms of connection and processing possibilities offered to the users and in terms of possibiities for plugging in different MT technologies. Work is organised along three action lines, i.e. data, MT engines and service/workflow[2].

- MT engines, i.e. the systems that carry out the translations for each language pair;
- MT data, i.e. the translation memories and language resources that each engine needs in order to produce its translations;
- MT service, i.e. the infrastructure that re-

[2]The first release of the full service is planned for the 2nd half of 2013.

ceives machine-translation requests and delivers completed translations to customers.

The action line on MT engines started by producing prototypical translation engines based on open-source SMT technology taken from ongoing reseach projects like EuroMatrix (Plus) co-funded by the European Commission, in particular the Moses platform (Koehn et al., 2007). Engines for 52 language pairs have been built and are currently (September 2011) available for internal testing. The training data for these engines was extracted from Euramis, the central translation memory of the EU institutions (Blatt, 1998).

These engines cover translation from English into all other official languages and vice versa, as well as a number of other frequently used combinations, like pairs involving French and German. The number of parallel segments that could be used for training ranged between 560 K for English - Irish to almost 14 M for English - French.

After a first quality assessment (see Section 4 for details), DGT selected ten language pairs with most promising performance and set up an automated process to pre-translate all incoming translation requests involving these language pairs and make resulting machine translations available to translators in form of TMX files. For the remaining language pairs, a machine translation can be requested manually at any moment by DGT staff through a test interface.

Ongoing improvement work on all the language pairs will allow DGT to automate pre-processing of all requests involving English, either as source or target language, in the context of the real-life trial by the end of the year.

2 Translation memory (TM) technology in the European Commission: Euramis

The Euramis project (European advanced multilingual information system) was launched in 1995 following a call for tenders for the Development of multilingual tools and their integration into multilingual services. The underlying idea was to relieve translators of the more repetitive work and to achieve greater consistency in language and methodology, thus contributing to better quality assurance.

Euramis is based on a client-server architecture and can now be accessed by users in the Council, the Court of Auditors, the Court of Justice, the Committee of the Regions, the European Economic and Social Committee, the Parliament and the Translation Centre for the Bodies of the European Union. This improves consistency and allows genuine data sharing between translators working for different EU institutions and bodies. Quality assurance is a major concern all the institutions that make use of Euramis.

The Euramis central translation memory is not used directly during the translation process: it is merely a database layer which is accessed to retrieve or store data processed locally by the translators, using Computer assisted Translation tools and text editors as front ends. At present, the Euramis central memory contains more than 450 million segments in all official EU languages. Automated Euramis pre-processing (retrieval) is carried out on all original documents in Word, HTML and XML format.

3 SMT as input to human translation

From an abstract point of view, translation memory lookup and SMT can be regarded as different ways towards the same goal. Both approaches are based on the assumption that the re-use of existing translations can be beneficial in the creation of new translations of similar documents. TMs recycle translations of complete sentences, which leads to high accuracy, but limited recall. In the case of exact matches, translations tend to be correct even in the new context. However only a small fraction of the segments lead to exact matches, whereas most cases have to resort to fuzzy matches or need to be translated from scratch. In contrast to this, SMT can recycle translations of even small snippets (so called "phrases") of text and re-combine them in novel ways so that complete input sentences can be covered. This dramatically increases the recall, so that every input sentence will be translated somehow, but due to inherent difficulties with ambiguity, word-aligment errors, and the lack of linguistic knowledge for re-combining the phrases in syntactically well-formed ways, the result is often of low accuracy.

Anyhow, the similarities between the two approaches are sufficient to employ existing tools for TM-based computer-aided translation (CAT) also in the case of SMT or to combine the two approaches. From the perspective of the translator, the workflow is fairly similar: In addition to the TM lookup results, the translator can ask for SMT of the source document to be translated. The result can optionally be delivered in the form of a TMX file, so that it can be used exactly in the same way as the TM retrieval results. It is also possible to combine TM retrieval with SMT by loading multiple TMX files into the translator's work bench. It is possible to assign penalties to the SMT results in order to determine which of the two will be shown to the translator first[3]. Clearly, the the penalty should be set in such a way that an exact match from TM will take priority over an SMT result, whereas SMT results are still preferred over fuzzy matches of a very low score. The exact threshold depends on many factors, such as the typical quality of SMT, the type of the document, but also on the personal preference of the translator. Confidence scores that could be computed for the SMT output, e.g. following the techniques proposed in (Specia et al., 2009), might be beneficial to streamline this presentation by calibrating the confidence score against TM match rates.

These simple ways to combine TM with SMT allow the selection of translations on the coarse granularity of complete segments. In the longer run, better tools for CAT should be developed that allow translators to make best use of the options contained within the search space of an SMT decoder or any hybrid MT system, and the Caitra tool (Koehn, 2009) shows some interesting possibilities, where the effect on translation speed and quality for ten translators from an academic context has been reported in (Koehn and Haddow, 2009). However, in order to be adopted in a professional translation context, such novel tools need to find a fine balance by showing a sufficient number of alternative expressions to the translators without overstraining their attention. The ergonomic and engineering challenges of building such tools that are mature for real-life application have not yet been fully met.

[3]The other translation can also be inspected, however at the price of additional mouse clicks.

4 MT quality control and evaluation

An effort to create a large number of MT engines and make them available to translators in a CAT context requires meaningful ways to assess the utility for the translators. Sufficiently high quality provided by MT should ultimately reduce the effort required on the side of the translator, increase the quality of the final product and might even improve the conditions of work for the translators, who could focus on the more creative aspects of the work, whereas the machine would free them of the more tedious parts (Kay, 1997).

However, before achieving such ambitious goals, we need to establish practical ways to measure progress in this direction.

A widespread way to measure the performance of MT systems is based on BLEU scores (Papineni et al., 2002). While BLEU scores have weaknesses and are not suitable for ranking MT systems accross different technologies (Callison-Burch et al., 2006), they are frequently used for tuning the parameters of a given SMT engine, and also our project makes systematic, yet implicit use of BLEU scores via the minimum error rate training (MERT) approach that is part of the Moses toolkit (Koehn et al., 2007).

Whereas BLEU scores can help to assess relative improvements between various versions of an MT engine, they do not appear suitable as a sole basis for decisions about the overall usability of an MT engine for any particular purpose. The criteria that should ultimately be applied in a CAT context are whether MT achieves improvements related to productivity, quality, and work conditions as mentioned above. These improvements, however, are not easy to measure, especially while the system is still in an early phase of its development. Even if SMT nowadays often shows rather promising results when doing "inbound" translation for understanding the gist of foreign language documents, it is very far from obvious that its use as a CAT tool in an institution creating large volumes of "outbound" translations into morphologically complex languages, which are often published as legally binding documents and hence need to be flawless, will be accepted by the translators. On the other hand, we received very encouraging early feed-back from some prospective users telling us that the translations into certain languages, although based on our very first prototypes, might already provide some benefit to translators who know how to deal with the imperfections of MT and who have already collected experience in combining results from MT with TM in the daily practice[4].

DGT therefore decided to make a first round of tests to find out about the usability of the current MT performance for CAT and to identify the main issues related to MT quality across all languages and a wide set of domains for which the sytems is supposed to be used. This so-called *maturity check* took place in April and May 2011 and was performed by 61 translators from 21 language departments[5], who assessed the quality of translations from English into their respective target language by deciding, sentence by sentence out of several hundreds, whether the machine translated text would be useful, i.e. requiring acceptable editing effort to get something comparable to the human quality translation, or useless[6]. This effort yielded more than 16 thousands individual judgments split over 9 selected documents varying from 4 to 9 documents per target language. Statistics over these judgements are summarized in the chart in Fig. 1, together with some annotation grouping the languages into families and sketching the main difficulties for translation into these languages.

Based on these findings, it was then decided by the respective language departments that for ten of the target languages, SMT results should be provided in an automatic fashion and delivered to translators together with the results of TM lookup, as part of the so-called *real-life trial*.

5 Towards a combination of TM and SMT

In conjunction with the so-called *maturity check*, works in several directions are ongoing in order to improve the quality of our current MT engines. Particular efforts are made concerning the combination of TM and MT approaches. In a CAT environment, the TM-based approach consists in searching a huge database of translations of complete sentences (so called TM) for an exact or approximate match of a new sentence. On the contrary, following the SMT approach, small phrases are selected according to existing translations (so called *training data*) and re-combine in order to generate the translation of a new sentence. As already said, the use of TMs leads to high accuracy but limited recall whereas SMT engine always provides automatic translations but often of low accuracy. The idea of combining the two approaches comes naturally in order to achieve both high accuracy and high coverage.

Many methods existing in the literature aim to combine TM with an SMT system.
The simplest way is of course to produce the ouput by using either TM or MT depending on the sentence to translate. One can imagine for example to use MT when nor exact or approximate match is founded in the TM. Unfortunately, it would mean that in this case, the sentence is quite different from the existing translations and consequently harder to translate even for an SMT system. (Biçici and Dymetman, 2008) combined a phrase-based SMT system with the TM match by extracting a phrase table from the TM match and adding it to the initial phrase table of the system. Their experiments showed that integrating TM into an SMT system allows significant improvements in terms of BLEU and NIST scores over the stand-alone SMT system and the stand-alone TM system for a translation task from English into French. (Simard and Isabelle, 2009) proposed to integrate a phrase-based MT system within a CAT environment. They presented two combination strategies. The

[4] In particular, colleagues from the Portuguese language department had done some pioneering work by using a home-brew installation of Moses and propagating its use amongs DGT translators, including running several studies on perceived quality and user satisfaction, which showed very promising results.

[5] The maturity check for English→Irish is currently ongoing, as a model for this language pair became available only later.

[6] The definition of the boundary was left to the translators involved in this exercise, and considerable individual differences could be found in the judgements. The proposal of a more fine-grained distinction into several classes was discussed in the user group, which then however decided to stick to the binary distinction that appeared easier to handle.

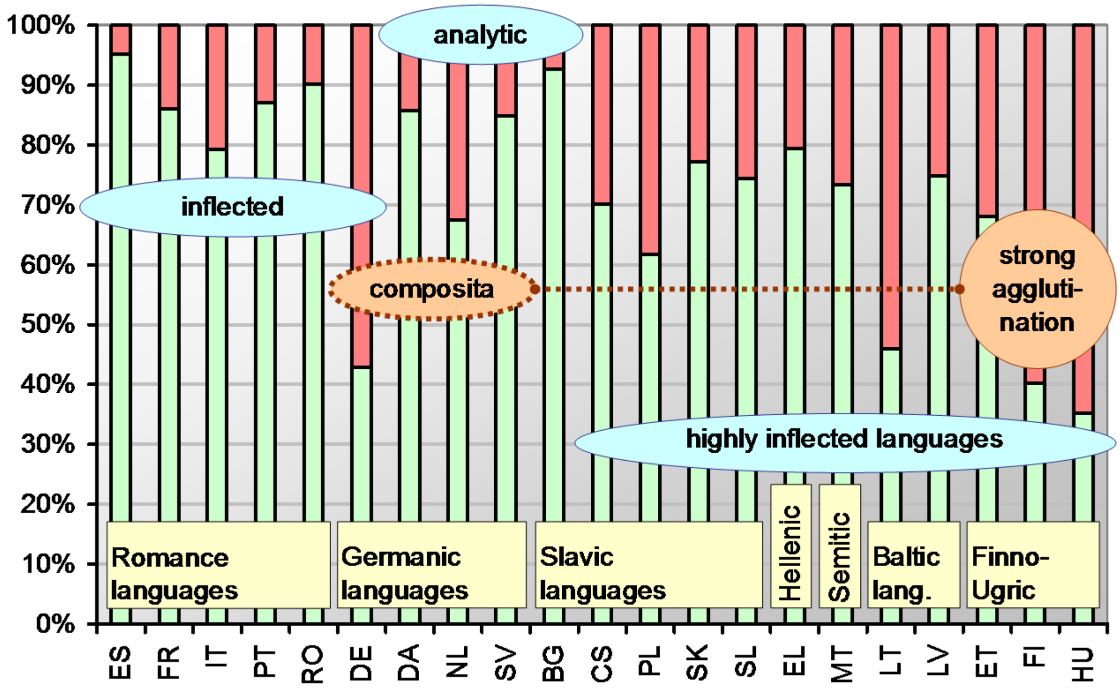

Figure 1: Outcome of a first "Maturity Check" for translations from English to other EU languages: Rate of segments labeled as useful, grouped by morphologic properties

first one consists in selecting the most appropriate approach (TM or MT) given the input sentence whereas the second one is very similar to the method of (Biçici and Dymetman, 2008) except that the TM-based phrase table does not allow the use of discontiguous phrases. Their experiments showed significant gains in MT quality for English-French when close or exact matches for the input sentence were found. More recently, (Koehn and Senellart, 2010) presented two methods which combine TM and SMT in different ways. In the first method, they proposed to construct XML frames that provide specified translations subsequences of the input sentence (the matched part) and rely in the SMT decoder to fill in the remainder (the unmatched part). The conducted experiments showed that this method outperforms TM and SMT for high fuzzy match ranges (80-99%). In the second method, these XML frames are encoded as very large hierarchical phrase rules, and used in a secondary rule table of a hierarchical phrasebased model. This method outperforms SMT even with lower fuzzy match ranges.

All these studies showed the advantages of combining the TM approach with the SMT approach in order to improve the translation quality. Given the large amount of TMs and the tools already implemented within the Euramis platform to access them (for different kind of operations such as: retrieve, storage, alignment ...), we therefore decided to implement one of these strategies. In the first place, we decided to follow an approach similar to the first one presented in (Koehn and Senellart, 2010), in which external knowledge from the TMs are provided to the SMT engine by adding XML markup in the source document to be translated. Indeed, as we currently have more than 50 SMT systems to manage that are already built and available for internal testing, it was more convenient for us to only act on the source document than to modify the models involved in all the engines.

When an SMT decoder has to translate a source sentence, it begins with an empty hypothese which it extends by using the entries of a large phrase table until generating a complete translation. For a given phrase or subsequence in a source sentence, the number of corresponding entries in the phrase table can be very big, which makes a correct selection quite difficult. By using the TMs, the idea is to determine a specific and correct translation for a given source phrase instead of letting the decoder search amongst all possibilities in the phrase table. To demonstrate the idea, we will use the following example taken from the data in the maturity check:

input:
```
The entity recognises a deferred
tax liability of 8 (40 at 20%) if
it expects to sell the itemwithout
further use and a deferred tax
liability of 12 (40 at 30%) if it
expects to retain the item and recover
its carrying amount through use.
```

output:
```
L'entité comptabilise un passif
d'impôt différé de 8 (40 à 20%).
S'il prévoit de vendre le poste sans
autre utilisation et un passif d'impôt
différé de 12 (40 à 30%) s'il s'attend
à de conserver le poste et de recouvrer
sa valeur comptable par l'usage.
```

ouput denotes the translation produced by our English-French engine without any external knowledge. We can notice some important errors like the introduction of a fullstop which induces the presence of two sentences in the final translation. This example indicates that the decoder does not always choose the appropriate entries in the phrase table. Now, if we look the existing translations in the TMs, we cand find a source sentence not exactly the same as the input sentence but very similar:

SRC:
```
The enterprise recognises a
deferred tax liability of 8 (40 at
20%) if it expects to sell the asset
without further use an a deferred
tax liability of 12 (40 at 30%) if it
expects to retainthe asset and recover
its carrying amount through use.
```

TGT:
```
L'entreprise comptabilise un passif
d'impôt différé de 8 (40 à 20%) si
elle s'attend à vendre l'actif et
ne plus l'utiliser, et un passif
d'impôt différé de 12 (40 à 30%) si
elle s'attend à conserver l'actif et à
recouvrer sa valeur comptable par son
utilisation.
```

We indicate in grey the words which differ between the input sentence and the source sentence found in the TMs. We can see that only the words "*enterprise*" and "*asset*" mismatch, therefore the idea is to let the decoder find the appropriate translation for "*entity*" and "*item*" but take the target side of the TM match as the translations of the common parts of both the *input* and the *source* sentences. To achieve that, it is possible to use an advanced feature within the Moses framework which allows the specification of translations for parts of a source sentence by using XML markup. We can simply tell the decoder what to use to translate certain words or phrases in the source sentence by surrounding them by tags. Given our example, the source sentence becomes:

```
<np translation="L'">The</np>
enterprise
<np translation="comptabilise un
passif d'impôt différé de 8 (40
à 20%) si elle s'attend à vendre
l'">recognises a deferred tax
liability of 8 (40 at 20%) if it
expects to sell the</np>
asset
<np translation="et ne plus
l'utiliser, et un passif d'impôt
différé de 12 (40 à 30%) si elle
s'attend à conserver l'">without
further use an a deferred tax
liability of 12 (40 at 30%) if it
expects to retain the</np>
asset
<np translation="et à recouvrer
sa valeur comptable par son
utilisation.">and recover its
carrying amount through use.</np>
```

and Moses produces the following automatic translation:

```
L' entreprise comptabilise un passif
d'impôt différé de 8 (40 à 20%) si
elle s'attend à vendre l' actif
et ne plus l'utiliser, et un passif
d'impôt différé de 12 (40 à 30%) si
elle s'attend à conserver l' actif et
à recouvrer sa valeur comptable par
son utilisation.
```

Given a new sentence *input* to be translated, the different steps of the general method are:

1. Find the best matching pair <SRC,TGT> for *input* in the training corpus. If it corresponds to an exact match of *input* then consider TGT as the final translation and go to step 4.
2. Identify the subsequences in *SRC* which exactly match with *input*.
3. Using the word-alignment previously computed on the training data, identify the corresponding subsequences in *TGT*.
4. Using the XML markup feature of Moses, add the identified subsequence pairs as external knowledge. Therefore, it is possible to plug in these translations to Moses without affecting the models.
5. Translate *input* using Moses

In the next section, we discuss the experimental set-up as well as the obtained results.

6 Experiments

6.1 Corpora

We conducted experiments for 22 language pairs going from English to all the other EU languages. Our experimental data consist of parallel corpora extracted from the TMs of DGT. For all language pairs, the experiments were performed according to three data sets: the training corpus used to build the system, the tuning corpus used to optimize the parameters involved in the decoding process and the test set used to evaluate the quality of the SMT systems and to compare them.
The training corpus consists of the TMs saved at DGT until October 2010. The table 1 reports the statistic of all training data in millions of sentence pairs. Each tuning corpus is made of 1000 sentence pairs extracted from the TMs saved between October 2010 and February 2011. Finally, for each language pair, we pick randomly a set of documents translated and saved at DGT after April 2011 to constitue the test corpus. The statistics of the different test sets are reported in table 2.

lang.	#training	lang.	#training
BG	5.3	IT	7.8
CS	4.5	LT	4.9
DA	7.8	LV	5.2
DE	9.9	MT	4.4
EL	6.3	NL	8.3
ES	8.9	PL	5.5
ET	5.2	PT	8.7
FI	5.6	RO	5.4
FR	13.9	SK	5.7
GA	0.56	SL	5.3
HU	5.8	SV	7.5

Table 1: Size of the different parallel corpora used for training (in millions of sentence pairs, ISO 639-1 codes, with English as the other language)

lang.	#sentences	lang.	#sentences
BG	475	IT	723
CS	745	LT	736
DA	324	LV	610
DE	637	MT	651
EL	644	NL	693
ES	627	PL	788
ET	785	PT	711
FI	615	RO	891
FR	499	SK	843
GA	–	SL	776
HU	723	SV	561

Table 2: Size of the different test sets used for evaluation

6.2 System configuration

6.2.1 Baseline Systems

The MT systems for the 22 language pairs were trained and tuned in the same way using the Moses toolkit. We used a baseline set-up as indicated during the WMT evaluation campaigns with lexicalized reordering and 5-gram language models. The parameter optimization was performed using minimum error-rate training (Och, 2003) on

BLEU score.

6.2.2 Preprocessing of the source sentence

The method we adopted in order to combine TMs with SMT does not require to change the baseline MT systems. We just have to preprocess the source document adding XML markup.
The first step is to retrieve the best match from the TMs according to a distance also called fuzzy match score. To achieve that, we used functionality for TM-RETRIEVAL already put in place within the Euramis system. This function allows to identify similar sentences from the TMs by using a fuzzy key (FK) mechanism. A FK is a numerical representation of a given sentence based on all sequences of three letters in the sentence. This representation is not sensitive to numbers, special characters, lower and upper case and permutations of the words within the sentence. The FK is automatically created for each sentence saved in the TMs and a specific function based on the Levenshtein distance is used to compute the similarity between two FKs and thus to retrieve similar sentences to a given input. As well as finding the best match of an input sentence from the TMs, the TM-RETRIEVAL function allows also to identify the mismatch part between the input and the source segment of the TMs. For our experiments, we set the fuzzy rate to 80% which means that the retreived sentence from the TM must have at least a matching rate of 80% compared to the input sentence. If no sentences in the TMs achieves this rate, the TM-RETRIEVAL function returns an empty sentence.
Once we have found the best match from the TMs (i.e a source segment associated with a target one) and the mismatch between the input and the TM source segment, it is easy to locate the subsequences which are exact matches of the input. A more difficult problem is to identify the corresponding subsequences in the TM target segment. To achieve that, we use the word alignment previously computed during the training step of our engines with the GIZA++ tool (Och and Ney, 2003) which is part of the Moses framework. Indeed, as the training data used to build our engine are extracted from the TMs, the best match found in the previous step is inevitabily inclosed within. The word alignment then allows to identify in the subsequence pairs between the TM source ang target segments which are translation from one another.

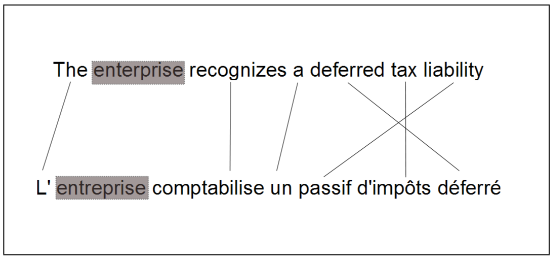

Figure 2: Word-alignment within a TM segment pair

Figure 2 shows the word alignment of one part of the TM segment pair based on our previous example. It leads to two translation pairs, that is to say: *”The → L’”* and *”recognizes a deffered tax liability → comptabilise un passif d’impôts différé”*
Finally, those translation pairs are inserted as external knowledge in the original document by using XML markup, like presented in section 5. Moses is then instructed to use this external knowledge in its output.
The results of our experiments are presented in the following section.

6.3 Results

We compare the method based on the combination of SMT with TM against our baseline engines. In order to not give the combined method an unfair advantage over the SMT baseline, only TM matches were used that come from the same years as the training data for the SMT models. The figure 6.3 shows the obtained results in terms of BLEU score. These results are preliminar and correlate only a partial implementation on what we have in mind. Indeed, for time and ressource reasons, we traited only simple cases. All complex cases (non-contiguous word alignment, unaligned matched words . . .) are still considered as no best match is found in the TMs.
The serie ”SMT” indicates the performance of our engine without adding any external knowlegde inside the test set whereas the serie ”SMT+TM” corresponds to the performance of our engines when external knowlegde extracted from TMs are added in the input text.

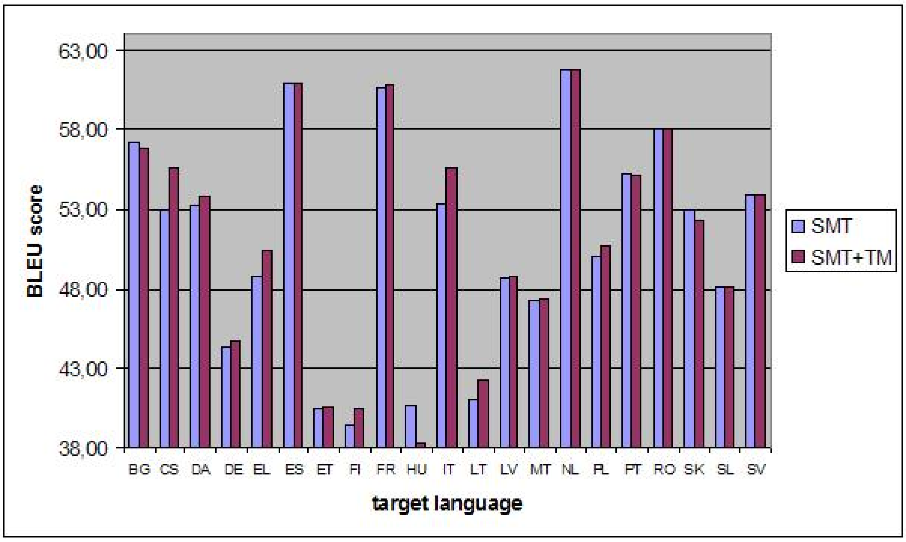

Figure 3: Bleu Scores of SMT compared to TM+SMT.

Except for the English-Hungarian language pair, adding external knowlegde from TMs gives similar or even better BLEU score compared to the use of pure SMT systems. Even considering only the SMT results, we can notice that BLEU scores are very high. That can be explained by the redundancy existing in the documents of DGT, and that is a reason why the use of SMT systems is very useful in our case. Nevertheless, significant improvments are obtained for translation from English into Czech, Greek, Italian, Finnish, and Latvian.

Even if it is difficult to draw solid conclusions about those results due to the small test sets we translated, we are convinced that combining SMT with TMs is a very promising solution in order to provide good MT documents to the translators. It Is also a good way to easily integrate their feedbacks in the SMT systems since TMs are updated daily with human translations.

7 Conclusion

In this paper we presented some of the activities carried out in the context of the MT@EC project of the European Commission, which has so far produced more than 50 baseline SMT engines, ten of which have already been integrated into DGT's translation workflow on an experimental basis, as part of a real-life trial. We presented ongoing work towards the more fine-grained integration of SMT with TM lookup, which has the potential to further increase the MT quality and make the overall system more useful both for translators and end users.

8 Acknowledgements

The work described in this paper was made possible via generous help from many colleagues within the IT unit of DGT and especially the team working on Euramis, as well as via the contributions of DGT's translators, either through the MT user group or indirectly through high-quality translations that were used for training the MT engines. We also thank anonymous reviewers for constructive comments on a draft version of the paper.

References

Biçici, Ergun and Marc Dymetman. 2008. Dynamic translation memory: Using statistical machine translation to improve translation memory fuzzy matches. Lecture Notes in Computer Science.

Blatt, Achim. 1998. Workflow using linguistic technology at the translation service of the european commission. In *Proceedings of the third EAMT*

Workshop on translation technology: integration in the workflow environment, pages 7–18.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL 2006)*, pages 249–256, Trento, Italy, April. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F. Zaidan, editors. 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, July.

Kay, Martin. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23. First appeared as a Xerox PARC Working paper in 1980.

Koehn, Philipp and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of MT Summit XII*, Ottawa, Canada.

Koehn, Philipp and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *AMTA Workshop on MT Research and the Translation Industry*.

Koehn, Philipp, Joel Martin, Rada Mihalcea, Christof Monz, and Ted Pedersen, editors. 2005. *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics, Ann Arbor, Michigan, June.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*.

Koehn, Philipp. 2009. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore, August. Association for Computational Linguistics.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Simard, Michel and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. In *Proceedings of the twelfth Machine Translation Summit*, pages 120–127, Ottawa, Ontario, Canada, August.

Specia, Lucia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proc. of the 13th Annual Conf. of the European Association of MT, EAMT09*, pages 28–35.