LetsMT!

Platform for Online Sharing of Training Data and Building User Tailored MT

www.letsmt.eu/

Project no. 250456

D6.3 Automatic evaluation report of domain specific SMT systems

Version No. 2.0

31/10/2011

Document Information

Deliverable number:	D6.3
Deliverable title:	Automatic evaluation report of domain specific SMT systems
Due date of deliverable according to DoW:	31/10/2011
Actual submission date of deliverable:	31/10/2011
Main Author(s):	UCPH: Lene Offersgaard
Participants:	UCPH, TILDE
Reviewer	MOR
Workpackage:	WP6
Workpackage title:	MT usage in localisation: facilities and evaluation
Workpackage leader:	MOR
Dissemination Level:	PU
Version:	V1.0
Keywords:	Evaluation, domain, SMT, localisation, BLEU, NIST, METEOR, TER

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval Level
0.8	23/08/2011	Draft	UCPH		Uploaded to project web site
0.9	26/08/2011	Draft	UCPH	Added evaluation results for en-pl-it system	Ready for partner review
1.0	31/08/2011	Final	UCPH	Review comments	Final, ready for submission
1.9	31/10/2011	Draft	UCPH	Updated results including new systems.	Ready for partner review
2.0	31/10/2011	Final	UCPH	Review comments	Final, ready for submission

EXECUTIVE SUMMARY

This document gives an overview of evaluation results for localisation SMT systems by M20(October 2011). The report will be updated twice during the development phase. This is the first update. Results from 4 systems are presented.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LetsMT!	Platform for Online Sharing of Training Data and Building User Tailored MT
API	Application Programming Interface
BLEU	BiLingual Evaluation Understudy
CAT	Computer aided translation
CRM	Customer relationship management
CSV	comma-separated values
ERP	Enterprise resource planning
GUI	graphical user interface
IPR	Intellectual property rights
Locale	Market with specific language, legal, cultural etc. needs. Locale is typically the same or smaller than a country, such as DE-DE or FR-CA, but can be also larger, such as ES-LA, which is rather a useful abstraction motivated by economies of scale than a real locale.
L10N	Localization - Creation of locale specific versions of products, documentation, and support materials. Translation is typically an important part of L10N process.
LSP	Language service provider
METEOR	Automatic Metric for MT Evaluation
MT	Machine translation
OLAP	Online analytical processing
SOV language	Languages with word order: Subject-Object-Verb
TBX	Term Base eXchange
TDA	TAUS Data Association
TER	Translation Edit Rate
TMX	Translation Memory eXchange format
TM	Translation memory
XLIFF	XML Localisation Interchange File Format

1 Introduction

This report documents the evaluation work in task 6.3. The aim of this task is to evaluate the trained SMT systems covering the localization case using automatic metrics. These results will allow to track incremental improvements of the systems and to highlight areas for improvements.

This report is closely connected to D3.6 "*Training and evaluation of initial SMT systems"*, which describes the chosen evaluation measures in detail and discusses the pros and cons of the used automatic evaluation measures.

This report will be updated twice during the project period by adding new evaluation results and other findings. This version is the first update. As the first deadline for trained localisation systems is M20 (Oct. 2011), the first version of the report only contained systems already trained and ready for testing by M18. The report is now updated by M20 presenting results from 4 systems.

2 Initial domain specific SMT systems

The initial domain specific SMT systems are trained as described in D6.2 *"SMT systems trained on domain specific data for usage in CAT tools"*. Please note that systems covering the specific domain: business and financial news domain are covered in work package 5 and are described in D5.3 and evaluated in D5.4.

The training data are available in the Resource Repository. The systems trained so far as part of work package 6 is:

- English –> Latvian IT: Information technology and data processing (short name: en-lv-it)
    - Version M18 and version M20
- English –> Polish IT: Information technology and data processing (short name: en-pl-it)
    - Version M18
- English –> Lithuanian IT: Information technology and data processing (short name: en-lt-it)

  Version M20

For each language combination several versions will be trained during the project period, where different selections of training data will be used. For the results in this delivery, the English-Latvian It system is trained in two versions with diffent amounts of training data. All the systems will be trained on the currently available in-domain parallel training data, with a selected combination of additional parallel and monolingual data. Details about the training process and the systems can be found in D6.2.

This report will focus on the automatic evaluation results for the systems trained by end of October 2011.

2.1 Evaluation sets

When evaluating SMT systems by means of automatic measures it is necessary to have evaluation corpora consisting of text in the source language with at least one corresponding reference translation. This will in the following be called an evaluation set.

For the validity of the test, it is also important that the evaluation set consists of so-called "un-seen" text, i.e. text that is not included in the training corpus. Therefore, the evaluation set is extracted from the available data material before training and excluded from the training corpus.

Evaluation sets for the initial automatic evaluation are randomly extracted from the in-domain corpus for business and finance domain. For each language pair, the size of the evaluation set is 1000 segments. A segment can be a sentence or another text segment. The size of the test set is chosen to keep the number small enough to ensure the possibility for manual inspection, and still be a representative size of a test corpus. Furthermore we prefer to have the same size for all evaluation sets.

For each trained system the evaluation set is automatic extracted, and therefore different versions of each system will have different evaluation sets.

From M16 to M20 the system platform functionality has been enhanced. One of the enhancements now ensures that the evaluation data are selected from in-domain data. This leads to more reliable evaluation results, as the evaluation sets earlier had a potential risk to contain segments from general corpora or non-in-domain data. By a manual inspection of the evaluation sets, the evaluation results in version 1.0 of this report do not seem to be affected of this change.

2.2 Tuning sets

In addition to the evaluation set, a so-called tuning set is also separated from the amount of training material. The tuning set is used during the training process as a special tuning corpus for adjusting the translation models and thereby optimizing the resemblance of the generated translation output with the target language part of the tuning set. An automatic evaluation measure is also used during this optimization, and for translation systems based on the Moses translation system the most widely-used measure is BLEU. Note that this tuning process serves the additional purpose of optimizing the system to resemble translations close to those found in the tuning set. It is important to ensure that the text sections that are extracted for the evaluation set and the tuning set do not overlap.

The results stated in this report focus on evaluation results based only on evaluation sets. Results of translated training material or tuning sets are not presented. All tuning sets have the size of 2000 segments.

2.3 Challenges and quality of evaluation data

When measuring translation quality by means of automatic measures, the evaluation is (in general) based on comparing the translation output with one or more reference translations.

If the evaluation will have to be based on more than one reference translation, the source text will have to be translated by professional translators to produce these references. In LetsMT! we have decided to keep the automatic evaluation as simple and cost efficient as possible. Therefore the evaluations are based on only one reference which is the target language part of the evaluation set.

Since the evaluation set is extracted randomly and automatically, it is possible that pairs of sentences are only approximately parallel or badly aligned. The presence of such challenging sentence pairs in the evaluation set will certainly make it much more difficult to get a good evaluation result.

3 Short description of evaluation metrics

Detailed descriptions of the evaluation measures can be found in D3.6 "*Training and evaluation of initial SMT systems"*. We use automatic metrics, which are faster, simpler and less expensive than human evaluation. However, these measures have a number of weaknesses compared to trained human evaluators.

The automatic metrics used are:

- NIST
- METEOR

- BLEU
- TER

4 Initial evaluation results

The evaluation results for the available system with measures used so far can be seen in table 1.

Localisation	System name	BLEU	NIST	METEOR	TER
English-Latvian IT, Ver1(M18)	en-lv-it-v1	0.497	8.097	0.429	56.5
English-Latvian IT, Ver2(M20)	en-lv-it-v2	0.564	8.690	**0.490**	**52.1**
English-Polish IT, Ver1(M18)	en-pl-it-v1	**0.605**	**9.119**	0.353	75.9
English-Lithuanian IT, Ver1(M20)	en-lt-it-v1	0.147	2.531	0.106	116.1

Table 1. The results of the initial systems for the automatic metrics BLEU, NIST, METEOR, TER. BLEU and NIST figures (Case Sensitive scoring) can also be seen at https://letsmt.eu/Systems.aspx.

4.1 BLEU and NIST results

The BLEU scores for the 4 systems range from 0.147 for English-Lithuanian IT to 0.605 for English-Polish.

The BLEU figures below 0.30 often indicate very low translation quality, whereas BLEU figures above 0.50 indicate a translation quality that can be useful for post-editing. The NIST scores are correlated to the BLEU scores.

The results in table 1 indicate that the English-Lithuanian system has a very low translation quality. A manual inspection of the evaluation set reveals – without any knowledge of Lithuanian that there are alignment errors certain sections of the evaluation set.

The English-Latvian IT system has been retrained and the already good results for version 1 has been improved by version 2.

The English-Polish system has the best BLEU/NIST scores, and this system has therefore not been retrained.

4.2 METEOR results

The METEOR results are calculated for all systems. We have used version 1.2; however it is a stripped version, where only the module exact is included in the scoring. The weights are set to default values[1].

The best result for English-Latvian is 0.49 and for English-Polish 0.35. The English-Lithuanian system score is 0.10, still indicating that the system is very poor. The score for English-Polish indicates translation of medium quality. While the English-Latvian system seems to be better using

[1] Parameter values: -p '0.5 1.0 1.0' are claimed to behave well for a wide range of languages.

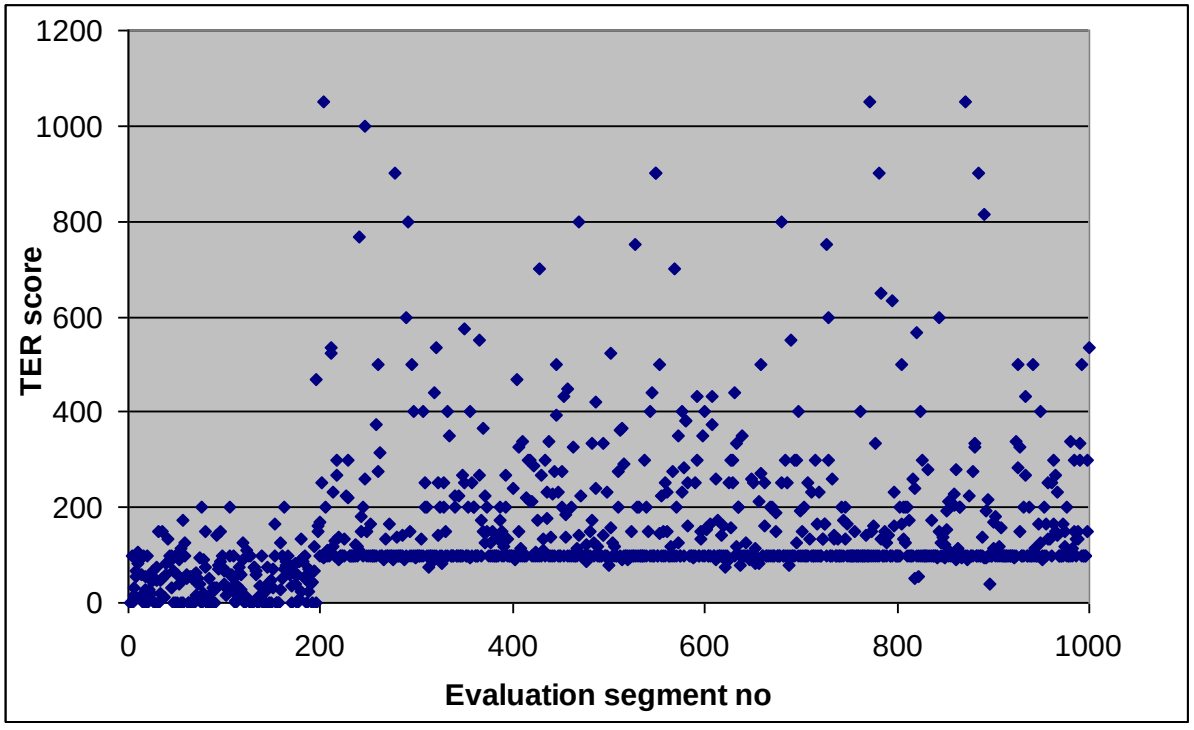
this score. These METEOR results are ranking these two systems differently compared to BLEU/NIST scores.

4.3 TER results

TER measures the number of insertions, deletions, substitutions and shifts in and compares this to the number of words in the sentence. Therefore a low TER score is better than a high score. The figures are given in %.

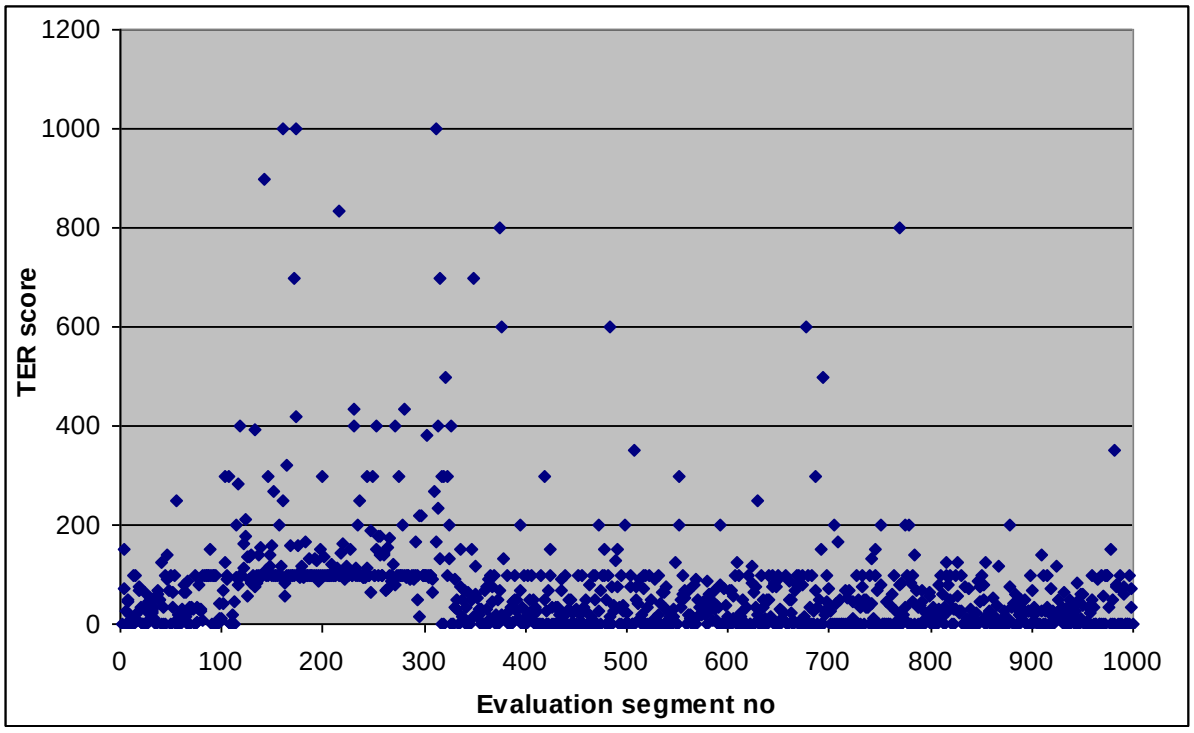
The English-Latvian v2 system gets the best score 52.1% compared to first version with 56.5%. Here we see the same improvement from version 1 to version 2 as for the other scores.
The English-Polish system has a score of 75.9%. This result agrees with METEOR ranking of the systems, stating the English-Latvian system as the system with the best performance.

The results for the English-Lithuanian system are very poor. All measures agree about this. In Figure 1 the TER scores for each evaluation segment are shown. Please mark that the first approx. 200 segments have reasonable scores, also a lot of TER scores close to 0% (0% ~ reference and translation are equal). The scores for the following show a large diversity and only a few of them are below 100%. This means that there is almost no similarity between words in reference and translation, and therefore indicating that the translation is very bad, perhaps caused by wrong alignments in data.



Figur 1. TER scores for all 1000 evaluation segments for the English-Lithuanian IT system.

For comparison the TER scores for the English-Latvian IT Ver2 system is shown in figure 2. Here only segments from approx. no 110 to approx. no 300 show systematic lesser TER score that the rest of the evaluation set. Clues to get a better automatic performance results for this system can perhaps be found in the data from those corpora represented in this part of the evaluation set.



Figur 2. TER scores for all 1000 evaluation segments for the English-Latvian IT Ver2 system.

4.4 Assessment of evaluation results

The evaluation results for the initial systems indicate that the systems for English-Polish and English-Latvian of medium quality useful for post-editing.

However, the automatic metrics do not rank the systems equally. Best score for each metric is marked with bold. This might be caused by intrinsic differences among the target languages (e.g. in word order).

It would be interesting to follow up on the automatic evaluation by some human evaluation of the translation output quality to prove that the system output is of medium quality, useful for post-editing.

We would expect the English-Latvian system to perform better in user test with a better translation quality than the English-Polish system, as we expect the METEOR and TER measures to weight the user important aspects higher than the BLEU/NIST measures.

The results in table 1 indicate that the English-Lithuanian system has a very, very low translation quality. A manual inspection of the evaluation set reveals – without any knowledge of Lithuanian that there are many alignment errors in certain sections of the evaluation set. The very poor results for this system might therefore be connected with the alignment quality of some of the corpora in the training data.

4.5 Amount of training data

All systems are trained with reasonable large in-domain resources. Adding more data might not improve the evaluation results, as the amount is around 1.5 mill sentences or much larger. Adding

more data improved the metrics for English-Latvian. For English-Polish more in-domain data might be a good option, but will not necessary improve scores for the system. For English-Lithuanian a system based on more in-domain data compared to general data and out-of-domain data might improve the scores, alternatively the quality of alignments in the corpora might be of larger importance.

System combination	Parallel training data (sentences)	Mono training data (sentences)	BLEU scores	METEOR	TER
English-Latvian IT, Ver1(M18)	1,780,539	1,887,339	0.497	0.429	56.5
English-Latvian IT, Ver2(M20)	7,155,296	42,035,335	0.564	**0.490**	**52.1**
English-Polish IT, Ver1(M18)	1,380,430	1,272,496	**0.605**	0.353	75.9
English-Lithuanian IT, Ver1(M20)	5,472,218	29,774,108	0.147	0.106	116.1

Table 3. Amount of training data together with the automatic scores.

5 Conclusion and recommendations for the next version of systems

In this section we will give the initial recommendations based on the results reported and we will list subjects for future work on evaluation.

Languages covered
The report presents the initial evaluation results for the four initial systems, covering three language pairs. According to the DoW this task evaluates systems trained as documented in D6.2. The list of language pairs will be extended during the project and evaluation results produced for these systems will then be reported. Training of English-Estonian It system was also initiated but the needed training time showed up to be too long to finish by the deadline for this report. The results for that system will be included in the next version of the delivery.

Amount of data
Training new systems with more data might be an option, but more appropriate is human evaluation or practical use of current systems.

Systems suitability for use of Support Group
A small human evaluation task carried out by partners will hopefully show that both English-Latvian IT, Ver2(M20) and English-Polish IT, Ver1(M18) are ready for use in Support Group.

Improvements
For English-Lithuanian thorough investigation of the data alignment quality should be carried out. Also a system using less data – prioritizing in-domain data – should be trained.

Language specific issues that might indicate a more complicated situation for SMT quality as rich morphology should also be taken into considerations before expecting the same performance across languages.

6 References

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.9416&rep=rep1&type=pdf

NIST 2005. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics". Retrieved 2010-04-17. Machine Translation Evaluation Official Results.
http://www.itl.nist.gov/iad/mig//tests/mt/doc/ngram-study.pdf

Snover, M., Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.
http://www.cs.umd.edu/~snover/tercom/

Lavie, A and Denkowski, M."The METEOR Metric for Automatic Evaluation of Machine Translation", Machine Translation, 2010
http://www.cs.cmu.edu/~alavie/METEOR/pdf/meteor-mtj-2009.pdf

Offersgaard, L., Povlsen, C., Almsteen, L., Maegaard, B., Domain specific MT in use, 12th EAMT conference, 22-23 September 2008, Hamburg, Germany
http://www.mt-archive.info/EAMT-2008-Offersgaard.pdf