LetsMT!

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

www.letsmt.eu/

Project no. 250456

D3.6 Initial SMT systems trained and evaluated

Version No. 1.0

31/08/2011

Document Information

Deliverable number:	D3.6
Deliverable title:	Initial SMT systems trained and evaluated
Due date of deliverable according to DoW:	31/8/2011
Actual submission date of deliverable:	31/8/2011
Main Author(s):	UCPH: Lene Offersgaard, Jürgen Wedekind
Participants:	UCPH: Lene Offersgaard, Jürgen Wedekind
Reviewer	UEDIN
Workpackage:	WP3
Workpackage title:	SMT training facilities and SMT web service
Workpackage leader:	UEDIN
Dissemination Level:	PU
Version:	V1.0
Keywords:	Training, evaluation, SMT, localisation, BLEU, NIST, METEOR, TER

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval Level
0.8	26/08/2011	Draft	UCPH		Uploaded to project web site
1.0	30/08/2011	Final	UCPH	Review comments from UEDIN	Reviewed and uploaded to project website

EXECUTIVE SUMMARY

This document gives an overview of the initial systems trained and evaluation metrics used in the initial evaluation. The specific details of the evaluation results of the initial systems can primarily be found D5.4 and D6.3 and these deliveries will be updated during the development phase. In this report an overview of the evaluation results August 2011 is included.

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 describes the initial SMT systems which is trained and evaluated by M18 (August 2011).

The work in task 3.6 depends on work carried out in the tasks 3.1- 3.5, and therefore benefits from other deliverables especially the deliverables:

- D3.3 "*SMT training facilities ready for integration*"
- D3.5 "*SMT Multi-model repository ready for integration"*

The focus of this deliverable is therefore to give an overview of the initial trained systems and to describe the evaluation metrics used to evaluate these systems.

A complete overview of trained systems is given in section 2, also including different test systems. Financial systems trained are described in detail in D5.3 "*SMT systems trained for business and financial news translation*" and systems with focus on localization are described in detail in D6.2 "*SMT systems trained on domain specific data for usage in CAT tools*".

As it is very important to evaluate the translation quality of the trained systems, automatic evaluation will be carried out for all systems. Translation evaluation is not done for those systems which are only used for software test". Section 3 will therefore explain the automatic evaluation metrics that have been used and the pros and cons of automatic metrics and human evaluation.

Concrete automatic evaluation results for the systems are available in the deliverables:

- D5.4 "*Automatic evaluation report of business and financial news SMT*"
- D6.3 "*Automatic evaluation report of domain specific SMT systems*".

Please consider these two series of deliverables as closely connected to this deliverable. Section 4 describes methods for evaluation, and section 5 summarizes the initial evaluation results.

2 Overview of trained SMT systems

In figure 1 all publicly available systems are listed. Four subject domains are represented:

- **Finance**: 8 systems
- **Law**: 1 system
- **Information technology and data processing**: 2 systems[1]
- **Biotechnology and health: 1 system**
- **Test sy**stems, with s**pecial testing purposes:** 4 systems

[1] Subject domain for English-Polish IT is by acident given as other in figure 1, it should be "**Information technology and data processing"**

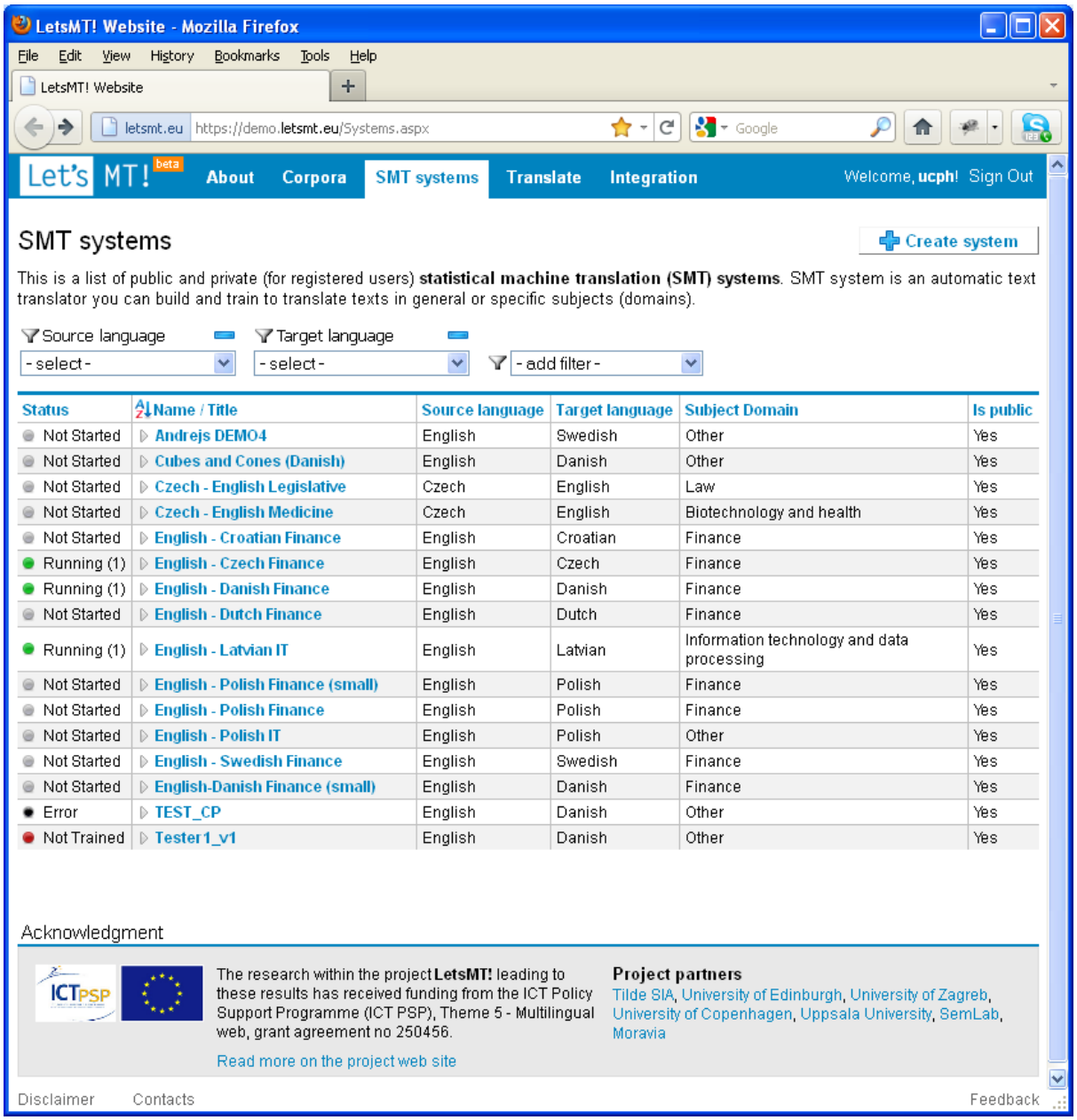

Figure 1: List of publicly available trained systems (August 2011)

Details about the systems can be found in D5.3 "*SMT systems trained for business and financial news translation*" and D6.2 "*SMT systems trained on domain specific data for usage in CAT tools*".

2.1 Evaluation data: quality and availability

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 sentences.

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 is 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 good evaluation results.

For data that are aligned within the LetsMT!-platform scores for the reliability of the alignments can be calculated, but for data already aligned when uploaded, it is difficult to realign and score the alignments, as these are primarily Translation Memories that are aligned by human translators and available in TMX format.

The evaluation results will therefore always depend on the evaluation set, resulting in very different scores using in-domain or more general evaluation sets. When you log into the LetsMT! platform you can find the evaluation and development sets can be found under the Edit tab on the LetsMT! homepage.

3 Description of evaluation metrics

The focus point of MT evaluation differs depending on your perspective. From a developers' point of view, evaluation has to be fast, simple and cheap. While from a users' point of view, the evaluation has to focus on east of use, better translation quality, quicker post-editing etc.

In LetsMT! we conducted a quick and cheap evaluation of all systems trained in the project. However, for some trained systems we were able to perform a more comprehensive evaluation. We mainly focus here on automatic metrics, but later in this section we will also describe more resource demanding evaluation possibilities.

In the following we provide a small overview of the evaluation metrics and describe their weaknesses compared to evaluation by trained human evaluators.

3.1.1 BLEU

The most widely used automatic metric for SMT is BLEU 'BiLingual Evaluation Understudy' (Papineni et al., 2002). Even though BLEU has been claimed to exhibit high correlation with human judgements, a number of weaknesses have been reported. The BLEU scores are weakly correlated to human evaluators on the sentence level, and even when BLEU results are given for a whole test corpus, the results are only in some cases proven to be correlated with human evaluators.

Calculations of scores are normally done for translated sentences by comparing them to a set of reference translations. The scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality.

BLEU results range from 0 to 1. The score indicates how similar the translation and the reference text are; values closer to 1 represent more similar texts.

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. These indications are based on the work in (Offersgaard, 2008) concerning Danish and English, but for other domains or languages with rich morphologies these approximated figures might not be useable. For the range 0.3-

0.5 we expect it to be unclear if the translation quality is useful for post-editing. It might depend on text type, language and subject domain.

3.1.2 NIST

NIST is a metric from the US National Institute of Standards and Technology. It is based on the BLEU metric, but with some alterations. Basically, BLEU/NIST metrics compare n-grams[2] of the candidate with the n-grams of the reference translation and count the number of matches. Where BLEU simply calculates n-gram precision assigning equal weight to each one, NIST also calculates how informative a particular n-gram is. That is, when a correct n-gram is found, the rarer that n-gram is, the more weight will be given to it (NIST 2005).
For example, if the bigram "on the" is correctly matched, it will receive lower weight than the correct matching of the bigram "interesting calculations", as this is less likely to occur.

The NIST scores are given as positive numbers, the larger the number the higher the similarity between the translation and the reference text. The maximum value of a NIST evaluation depends on the evaluation corpus.

3.1.3 METEOR

METEOR 'Metric for Evaluation of Translation with Explicit Ordering' (Lavie, 2010) is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stem and synonymy matching, along with the standard exact word matching. Therefore, language dependent resources (a stemmer and a synonymy resource) are required, which results in a more complicated setup process. The metric was designed to fix some of the problems found in the more popular BLEU metric.

The METEOR results are calculated for all systems using version 1.2; this is a stripped version, where only exact word matches are included in the scoring. This can be changed in the future, by allowing different word matching schemes. The weights are set to default values[3].

METEOR can also generate a number of analyses when performing evaluation. One of these is presented in figure 1 for English-Czech finance, where the score distribution for the number of individual sentences is given. These graphs will be more useful when comparing two systems, but it is included here to illustrate the distribution of the scores in the evaluation set. The figure shows that more than 200 sentences have a very low score (below 0.1). This might indicate that for some of these sentences the alignments are of bad quality.

[2] An n-gram is a sequence of any number of items (words) appearing in a document.

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

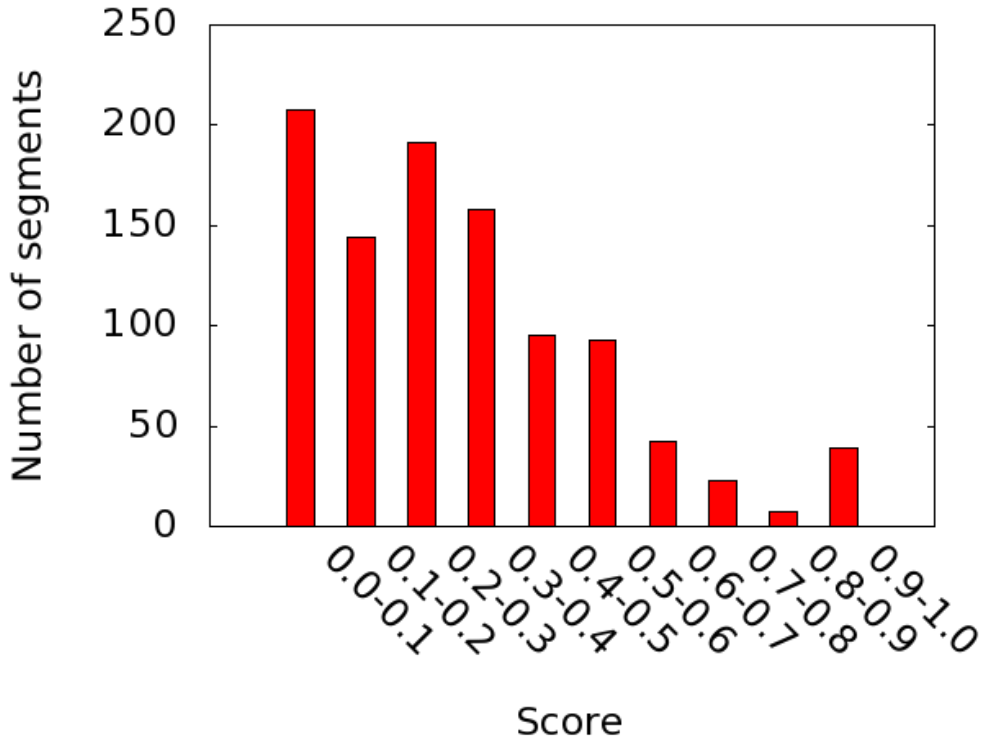

Figure 1. METEOR score distribution for sentences in the evaluation set for English-Czech.

3.1.4 TER

TER is an acronym for 'Translation Edit Rate' by (Snover et al. 2006). TER is an error metric for machine translation that measures the number of edits required to change the system translation into one of the references. TER is calculated as the count of insertions, deletions, substitutions and shifts of words divided with the number of words in the sentence.

As TER measures the number of corrections and compares this to the number of words in the sentence, a low TER score is better than a high score. TER is earlier stated to correlate reasonably well with human judgements (Snover et al. 2006).

TER values will be in the range from 0 (translated sentence is exactly like the reference) to in principle more that 100%, e.g. if the length of translation output is significantly longer than the reference translation then the number of edits might exceed the length of the reference translation.

3.1.5 TESLA

TESLA (Translation Evaluation of Sentences with Linear-programming-based Analysis (Dahlmeier et al. 2011) is a family of evaluation metrics which are based on n-gram matching but also take into account a varying degree of linguistic analysis. The simplest version, TESLA-M, utilizes lemmatization, POS tagging, and WordNet synonym relations. The more sophisticated variants, TESLA-B and TESLA-F, also exploit language models and a lightweight semantic representation of the target language obtained from the distribution over all phrase alignments of the target language sentences and their translations into a pivot language. Sentence-level scores for reference and system translations are obtained by evaluating similarity functions defined for the different linguistic features.

At WMT11 (Callison-Burch 2011) the TESLA metrics M and B demonstrated strong correlation with human judgments for the out of English direction. (Since the overall strongest metric MTeRater-Plus

relies on features specific to English, it has only been evaluated for translation into English.) Similar to METEOR, the TESLA metrics also depend on deeper linguistic information.

Thus evaluation can only be conducted for language pairs where the required resources are available.

3.2 Other evaluation methods

From a user's point of view, automatic evaluation figures are somewhat abstract and difficult to comprehend and do not necessarily provide feedback to the translator or user on the translation quality. Alternative evaluation metrics focusing much more on the human translation aspect have proposed to cope better with this problem. The following methods represent this alternative evaluation approach:

- Post-editing time
- Sentence ranking
- Fluency and adequacy scoring
- Usability scoring

All those methods require extended human interaction to perform the evaluation, and therefore we cannot include this kind of evaluation in the standard system development test. However, we found it interesting to test some systems later in the testing work in LetsMT! using some of these methods.

3.2.1 Post-editing time

When judging the profitability of using MT in a translation task, the most important factor is the effort needed to post-edit the MT output to the needed quality of the translation. When including MT in e.g. CAT tools, the effectiveness can be calculated by measuring the post-editing time with and with-out using MT. Thus, to produce such a comparison the same translation job has to be performed twice, using two different translators and that is a time-consuming task, depending on available translators.

3.2.2 Ranking of output sentences from different system

Another evaluation criteria used in the research community is ranking of sentences (Callison-Burch et al. 2011). This evaluation method is used for ranking comparable systems for the same language pair and inside the same domain. This evaluation method is not fitting our needs very well, as we primarily will have systems with small improvements from version to version, and not having all the versions available from the beginning.

3.2.3 Fluency and adequacy scoring

Fluency and adequacy have originally been defined using a five point scale (White 94). Later studies show that scores for fluency and adequacy apparently do not correlate very well between users, and therefore these score results are difficult to use as a reliable basis for system testing and tuning. Other studies (Offersgaard et al. 2008) have shown that the post-editors involved in evaluating SMT output stated that a five point scale would be much too difficult to use. If using fluency and adequacy measures we therefore suggest using only a four point scale which is easier to handle by the users and and probably more reliable. Fluency and adequacy scoring is less resource-demanding than post-editing, as the evaluator only has to do the judgement, not to write the correct translation.

3.2.4 Usability scoring

A simple scoring mechanism that has been suggested by a Danish LSP (Offersgaard et al. 2008) is usability scoring. 'Usability' is a measure that allows post-editors to score a machine-translated translation unit in terms of usability compared to a fuzzy match in a TM tool. A machine-translated translation unit may not be adequate or fluent, but it may be usable. When it is usable, the time needed

to edit the machine-translated translation unit will be shorter than the time needed to translate the segment from scratch. It is in this context defined as a three point scale. The scale is largely depending on the post-editing process, and the user can use the following scores:

> **3**: Good translation – few key strokes needed to edit translation. Corrections of casing or layout may be needed. Use of terminology is correct.
>
> **2**: Translation can be post-edited using less time than a translation of the sentence from scratch – number of key strokes needed to edit translation is less than the key strokes needed to translate from scratch.
>
> **1**: Translation quality is too poor. It will take more time to post-edit the sentence, than to translate the sentence from the source sentence – translation is discarded.

The post-editors at the LSP found this scoring very useful as it is closely connected to their translation workflow, no matter whether they use TM as their translation tool or post-edit MT-output.

3.3 Pros and cons of MT evaluation methods

Certainly, human evaluation is the undisputed gold standard for quantifying translation quality. But human evaluation is very time consuming. Moreover, objectivity and reproducibility is difficult to maintain. Automatic evaluation, on the other hand, is much cheaper and quicker than human evaluation and the scoring is objective. However, automatic measures have always been considered as a more or less crude approximation to the human assessment of translation quality.

Since BLEU, the most widely used automatic metric in MT, does not correlate especially well with human judgments, new metrics have been developed that aim to provide a stronger correlation. In most cases, however, they depend on language resources and tools that are usually available only for dominating languages.

Here, we included the traditional BLEU/NIST evaluation, because system results are usually reported using the BLEU score. We also considered TER as it can be calculated language independently and we used the basic version of METEOR, because it does not require language dependent resources.

The TESLA metric can be used in three variants, TESLA-M(minimal) depending on lemmatization, part-of-speech tagging and WordNet-lookup for target-language and TESLA-B(basic) depending on bilingual phrase-tables. TESLA-F (full) is the most sophisticated version. Initially we chose to leave out the TESLA score, as it requires complicated resources, which are not available for most of the target languages in LetsMT!

Experiments with TESLA-B might be carries out at later stages, because the required bilingual phrase-tables can be compiled during training phase.

4 Suggestions for future evaluation work

Chosen metrics

- BLEU/NIST
- TER
- METEOR

Human evaluation - in small scale

For systems where automatic evaluation indicates medium or good quality, human evaluation in small scale would be very interesting and appropriate.

This would both give information about the usefulness of BLEU/NIST, TER and METEOR for under-resourced languages that are in focus in the project, and give more information on which level automatic metric scores indicate 'usable' systems. Here the term 'usable' means translations either good or worth post-editing (scores 3 or 2 of the 'Usability' scoring above).

Amount of training data

The statement "More training data results in better systems" has often been heard in connection with SMT. Our hypothesis is that more in-domain data is good, but for in-domain systems, to much general data can lead to poorer performance. We will do some experiments with this.

Text types as an evaluation parameter

It would be interesting if we had enough data to train in-domain or general systems with different text types. Carrying out some evaluations of this would be very interesting, but could only be done if a lot of training resources for different text types are available, and this will therefore depend on the outcome of the data collection.

5 Initial evaluation results

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

System name	System name	BLEU	NIST	METEOR	TER
English- Czech Finance	en-cs-finance	0.346	7.349	0.308	59.7
English-Croatian Finance	en-hr-finance	0.219	5.850	0.198	73.4
English-Danish Finance	en-da-finance	0.275	6.076	0.249	72.5
English-Dutch Finance	en-nl-finance	0.222	6.114	0.215	72.6
English-Polish Finance	en-pl-finance	0.371	7.302	0.336	62.3
English-Swedish Finance	en-sv-finance	0.254	5.625	0.246	74.9
English-Latvian IT	en-lv-it	0.497	8.097	**0.429**	**56.5**
English-Polish IT	en-pl-it	**0.605**	**9.119**	0.353	75.9

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. The best score for each metric is in bold.

6 Summary

Domains covered

Four subject domains are represented in the initially trained systems August 2011: Finance, Law, Information technology and data processing, Biotechnology and health. The target at milestone M18 included three domains, so this goal is fulfilled.

Number of trained systems

14 systems are trained, eight can be seen in table 1, the other six are test systems. The target at milestone M18 included 10 trained systems, so this goal is fulfilled.

Evaluation methods

The following automatic metrics are chosen: BLEU/NIST, TER, METEOR. Including human evaluation in a small scale will be prioritised. The language pairs and domains will be decided later. The systems for human evaluation will be chosen by the systems with the best ranking based on automatic metrics.

7 References

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof and Zaidan, Omar, Findings of the 2011 Workshop on Statistical Machine Translation, SixthWorkshop on Statistical Machine Translation WMT11, July, 30–31, 2011, Conference on Empirical Methods on Natural Language Processing (EMNLP) 2011

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

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

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

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

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/

White, John and O'Connell, Theresa A.: Evaluation in the ARPA Marchine Translation, Program: 1993 Methodology. ACL (1994)

Dahlmeier, Daniel, Liu, Chang, Ng, Hwee Tou, *"TESLA at WMT 2011: Translation Evaluation and Tunable Metric"*. SixthWorkshop on Statistical Machine Translation WMT11, July, 30–31, 2011, Conference on Empirical Methods on Natural Language Processing (EMNLP) 2011