LetsMT!

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

www.letsmt.eu/

Project no. 250456

D5.4 Automatic evaluation report of business and financial news SMT

Version No. 1.0

30/06/2011

Document Information

Deliverable number:	D5.4
Deliverable title:	Automatic evaluation report of business and financial news SMT
Due date of deliverable according to DoW:	30/6/2011
Actual submission date of deliverable:	30/6/2011
Main Author(s):	UCPH: Lene Offersgaard, Jürgen Wedekind
Participants:	UCPH: Lene Offersgaard, Jürgen Wedekind, MOR: Tomas Hudik
Reviewer	SEM
Workpackage:	WP5
Workpackage title:	MT usage in news translation: facilities and evaluation
Workpackage leader:	SEM
Dissemination Level:	R
Version:	V1.0
Keywords:	Evaluation, domain, SMT, finance, BLEU, NIST, METEOR, TER

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/Approval Level**
0.9	27/06/2011	Draft	UCPH		Uploaded to project web site
1.0	30/06/2011	Final	UCPH		Uploaded to project web site

EXECUTIVE SUMMARY

This document gives an overview of initial evaluation results for business and financial news SMT systems. The report will be updated twice during the development phase.

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 initial evaluation work in task 5.4. The aim of this task is to evaluate the initially trained SMT systems covering the business and financial news case using automatic metrics. These initial 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.

2 Initial business and financial news SMT systems

The initial business and financial news SMT systems are trained as described in D5.3 *"SMT systems trained for business and financial news translation"*.

The training data are available in the Resource Repository. The outcome of the training process is a number of systems for the domain business and financial news:

- English –> Czech Finance (short name: en-cs-finance)
- English –> Croatian Finance (short name: en-hr-finance)
- English –> Danish Finance (short name: en-da-finance)
- English –> Dutch Finance (short name: en-nl-finance)
- English –> Polish Finance (short name: en-pl-finance)
- English –> Swedish Finance (short name: en-sv-finance)

For each language combination several versions will be trained during the project period, where different selections of training data will be used. For the initial results in this delivery, the systems will be trained on the currently available in-domain parallel training data, with different combinations of additional parallel and monolingual data. Details about the training process and the systems can be found in D5.3.

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

2.2 Development sets

In addition to the evaluation set, a so-called development set is also separated from the amount of training material. The development 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 development 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 development set. Therefore, additional results will be derived for optimized systems. It is important to ensure that the text sections that are extracted for the evaluation set and the development 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 development sets are not presented.

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 is based on more than one reference translation, the source text will have to be translated by professional translators to produce the 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"*. Here, only a brief description is given. We use automatic metrics, which are faster, simpler and less expensive. However, these measures have a number of weaknesses compared to trained human evaluators.

3.1 Evaluation metrics

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 is; values closer to 1 represent more similar texts.

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[1] 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.

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.

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.

4 Initial evaluation results

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

Business & Finance	System name	BLEU	NIST	METEOR	TER
English- Czech	en-cs-finance	**0.3464**	**7.3491**	**30.8%**	**59.7%**
English-Croatian	en-hr-finance	0.2194	5.8502	19.8%	73.4%
English-Danish	en-da-finance	0.2748	6.0758	24.9%	72.5%
English-Dutch	en-nl-finance	0.2221	6.1135	21.5%	72.6%
English-Polish	en-pl-finance	**0.3705**	**7.3019**	**33.6%**	**62.3%**
English-Swedish	en-sv-finance	0.2541	5.6247	24.6%	74.9%

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

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

4.1 BLEU and NIST results

The BLEU scores for the 6 systems range from 0.2194 for English-Croatian to 0.3705 for English-Polish. English-Czech has the second highest score: 0.3464.

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. The overall picture is that all the systems need to be improved to produce useful translations.

The NIST scores are correlated to the BLEU scores. Here the English-Polish and English-Czech also have the best scores.

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[2].

The result for English-Czech is 30.85% and for English-Polish 33.6%, which indicates a translation of medium quality.

For English-Croatian, English-Danish, English-Swedish and English-Dutch the scores indicate low translation quality.

METEOR generates a number of analyses when performing evaluation. One of these is presented in figure 1 for English-Czech, 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] Parameter values: -p '0.5 1.0 1.0' are claimed to behave well for a wide range of languages.

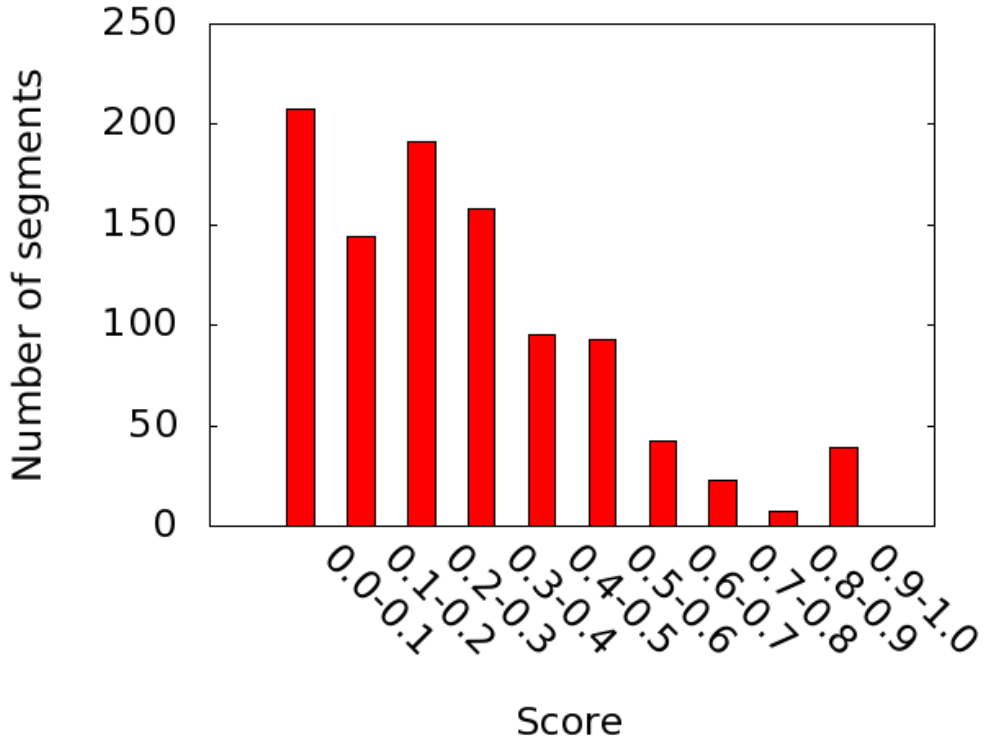

Figure 1. METEOR score distribution for sentences in the evaluation set for English-Czech.

The METEOR scores in this report are all derived with the exact module of METEOR. Later we will have to analyze to which extent we can include language specific resources for all target languages. For comparison purposes across languages it might be best to include the same number of modules in the METEOR calculation for all languages.

4.3 TER results

TER measures the number of insertions, deletions, substitutions and shifts and compares this to the number of words in the sentence. Therefore a low TER score is better than a high score. The English-Czech and English-Polish systems obtain the best scores for the TER measure. These results are in line with the other metrics.

4.4 Assessment of evaluation results

The evaluation results for the initial systems indicate that the systems for English-Danish, English-Swedish and English-Dutch need very strong improvements before they can be integrated into a web-site. The other three systems seem to be of better quality, but they also need strong improvements.

However, because of the intrinsic differences among the target languages (e.g. in word order) it cannot be expected that all systems will reach the same translation quality.

A brief investigation of the first part of the evaluation set and the translation output for English-Danish reveals that the reference translations are defective. Only a few examples are listed in table 2. It is important to mention that the Business & Financial texts are periodic reports including a lot of figures and tables, with documents in PDF file format.

No.	Source sentence	Reference Translation	Translation output
4	However, all current forecasts are subject to considerable uncertainty.	Alle prognoser er dog i denne tid Bloomberg), men væksten ventes at aftage i 4. kvartal. behæftet med betydelig usikkerhed.	Men alle nuværende prognoser er betydelig usikkerhed.
6	237 1,435 12%	167 237 1.435 12%	237 1.435 12%
9	part of the adjustment to the changed market conditions in the dry cargo market.	holdsvis 3 Handymax tørlastskibe og 1 Panamax tørlastskib.	del af tilpasningen til de ændrede markedsforhold i tørlastmarkedet.
989	Let us finally take Burundi 's problem seriously.	Lad os for en gangs skyld tage problemet i Burundi alvorligt.	Lad os tage Burundi.Jeg 's problem.Hr. alvorligt.
999	The European Union has provided emergency aid.	Den Europæiske Union har stillet midler til rådighed for den akutte indsats.	Den Europæiske Union har ydet nødhjælp.

Table 2. Examples of source and reference sentences from the English-Danish evaluation set compared with the system output.

It can be seen from these examples that some of the source-reference sentence pairs are not flawless. For sentence no. 4, the reference is badly aligned. The parts "Bloomberg)," and "væksten ventes at" is not relevant for the source sentence. The system output is not perfect but very close.

For sentence no. 6, the source is missing a number compared to the reference.

For sentence no. 9, the source and reference are not related. The problem could be that the reference does just not contain this sentence, but more likely that the alignment is bad, as the sentence starts with a lowercase word "part". The system output is a correct translation of the source sentence.

For sentence no. 989, source and target sentences are parallel, but here we can see some problems concerning the ".Jeg" and "Hr." in the output sentence. This error type should be investigated further to see if the system has difficulties to handle ".".

For sentence no.999, source and target covers the same meaning but is not strictly parallel. But the translation is perfect given the source.

These examples extracted from the evaluation set illustrate some of the problems connected to automatic evaluation of translation system. If the source and reference translations are not strictly parallel, the translation system has either an impossible or at least a very difficult job generating a suitable output compared to the reference.

We suggests that native speakers of the look though 10% of the sentences of the evaluation sets and check how well the sentences are aligned. If more that 10% of these are badly aligned, we should consider editing the evaluation set, replacing bad aligned sentences with well aligned sentences.

4.5 Amount of training data

Currently the systems are based on very different amounts of data. The smallest data amount is used for training the English-Croatian system: 59.000 sentences, while English-Dutch is based on the largest amount: See table 3. Given the facts that the English-Dutch and the English-Polish systems

have the best automatic scores with medium amount of training data, the conclusion might be that more data is not necessarily leading to better performance.

A more detailed analysis of the amounts of in-domain data versus general and out-of-domain data used training the systems has to be carried out before training the next versions of the systems. This analysis will lead to suggestions about the combination of training data.

System combination	Parallel training data (sentences)	BLEU scores	METEOR	TER
English- Czech	**536,680**	**0.3464**	**30.8%**	**59.7%**
English-Croatian	58,814	0.2194	19.8%	73.4%
English-Danish	1,353,572	0.2748	24.9%	72.5%
English-Dutch	1,505,706	0.2221	21.5%	72.6%
English-Polish	**742,136**	**0.3705**	**33.6%**	**62.3%**
English-Swedish	1,458,983	0.2541	24.6%	74.9%

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 initial systems. According to the DoW this task evaluates 6 systems trained as documented in D5.3, where the English-Slovak system mentioned in the DoW has been replaced by English-Swedish (for reasons mentioned in D5.3). The list of language pairs covered by the financial and business domain systems might be extended during the project and evaluation results produced for these systems will then be reported..

Evaluation sets for the domain Business & Finance
As described above we suggest to evaluate the evaluation sets by letting a native speaker randomly check for bad alignments in 10% of the evaluation sets, and if more that 10 % of the checked sentences are badly aligned then we suggest to clean up the whole evaluation set.

Amount of data
In general we can say that:
- if a SMT system should be developed for Croatian, more training data is needed
- for Danish, Dutch and Swedish an analysis has to be carried out to investigate if another selection of training data – even less training data – can give better performance.
- possibilities to add more in-domain training data should be investigated

New versions of systems should be trained before release of system
After the analysis of the selected training data for English-Danish, English-Dutch and English-Swedish has been carried out, a new series of systems should be trained. These new systems should be trained before announcing LetsMT! translation platform to the public or the Support Group.

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