LetsMT!

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

www.letsmt.eu/

Project no. 250456

D5.3 SMT systems trained for business and financial news translation

Version No. 3.3

31/08/2012

Document Information

Deliverable number:	D5.3
Deliverable title:	SMT systems trained for business and financial news translation
Due date of deliverable according to DoW:	31/08/2012
Actual submission date of deliverable:	31/08/2012
Main Author(s):	Tilde, SEM
Participants:	UCPH, SEM, UUP, MOR, FFZG
Reviewer	Tilde
Workpackage:	WP5
Workpackage title:	MT usage in news translation: facilities and evaluation
Workpackage leader:	SEM
Dissemination Level:	PU
Version:	V3.3
Keywords:	SMT training, business and finance

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval Level
0.1	21/10/2011	Initial Draft	SEM		Ready for comments
0.2	28/10/2011	Draft	Tilde	Document formatting, explanation of NIST/BLEU scores	Ready for comments
1.0	30/06/2011	Final	Tilde		Submitted
1.1	28/06/2012	Draft	SEM	Updated BLEU scores	Ready for comments
2.0	30/06/2012	Final	Tilde	Completely rewritten	Submitted
3.0	22/08/2012	Final	SEM	Added comparison with Google	Ready for comments
3.1	23/082012	PreFinal	Tilde	Reviewed	Ask for update
3.2	27/08/2012	PreFinal	SEM	editions	Ready for final review
3.3	31/08/2012	Final	Tilde	Reviewed, corrected	Ready for submission

Table of Contents

List of Tables

1 Introduction

This report is an updated version of the report submitted on M28. According to the reviewer's suggestion comparison with Google Translate was made. Summary of the comparison is included in the section "Benchmarking LetsMT!/Google Translate", detailed results are assessed in deliverable "D5.5 Evaluation report on usability of SMT in business and financial news translation" (M30).

This report describes the SMT systems trained and evaluated is scope of WP5 by M28 (June 2012). This is the third version of the report, the first version of the report was produced on M16 (June 2011) and it contained information about SMT systems trained by M16.

The aim of this deliverable is to provide a summary of SMT systems trained for the LetsMT! evaluation in business and financial news translation scenario. As it is important to evaluate the translation quality of the trained systems, automatic evaluation has been carried out for all trained SMT systems. Detailed results of automatic evaluation of SMT systems for the localization scenario is given in the deliverable D5.4 "Automatic evaluation report of business and financial news SMT" (M29).

Training of SMT systems was done by Tilde using the LetsMT! platform and business and financial domain specific corpora collected by SEM, UCPH and FFZG in scope of WP4. Corpora collected by SEM contained documents in PDF and DOC formats. These documents were automatically converted to text, aligned and converted to TMX format by UUP. An additional English-Danish training data cleaning and filtering was done by UCPH. System evaluation and tuning was done using manually verified and cleaned datasets. The manual verification and cleaning of evaluation and tuning sets was done by UCPH (Danish), SEM (Dutch), UUP (Swedish) and MOR (Czech and Polish). Tuning and evaluation sets for Croatian were manually translated by FFZG.

2 Trained SMT systems

Table 1 below summarizes all LetsMT! systems trained in Task 5.3 by M28. All these systems are in business and financial domain, as this is the domain used to evaluate LetsMT! platform in business and financial news translation.

This report summarizes information only about final trained systems (seven in total); many more experimental and testing systems have been trained in Task 5.3.

Table 1. List of final SMT systems for evaluation in business and financial news translation

Language pair	SMT System	BLEU score M20	BLEU score M28
English - Danish	English-Danish Finance IV	27.48	72.48
English - Dutch	English - Dutch Finance v3	22.21	62.98
English - Swedish	English - Swedish Finance v3	25.41	65.47
Dutch - English	Dutch - English Finance v3	56.18[1]	62.11
English - Czech	English - Czech Finance v3	34.64	59.91
English - Polish	English - Polish Finance v3	37.05	53.05
English - Croatian	English-Croatian Finance v3	21.94	17.42[2]

As it is shown in Table 1 translation quality has been significantly improved during last 8 project month. There are several reasons for such improvements:

1. Project partners have collected more parallel training data,
2. Project partners have collected more monolingual training data,
3. Evaluation and tuning sets have been manually verified and cleaned (in case of Croatian – manually translated),
4. Automatic text extraction from PDF and DOC files and automatic alignment has been improved,
5. Additional data filtering has been applied to filter out mistakes of automatic alignment;

Figure 1 and Figure 2 show the training phases and individual training steps for SMT system training. We trained SMT systems for business and financial news translation using two different approaches depending on the available training data. For language pairs where we had significant amount of

[1] 2012.04.10. System used for the initial usability evaluation of business and financial news SMT (D5.5)

[2] BLEU scores of two versions of Croatian systems cannot be compared because they have been calculated using completely different evaluation sets.

domain specific parallel training data (English-Dutch, English-Danish, English-Swedish, English-Polish) we trained systems using only in-domain data both for translation models and language models (Figure 1). But for languages where we did not have enough domain specific domain specific training data (English-Czech, English-Croatian), we trained (i) translation model using both in-domain and general data and (ii) two language models – one using general monolingual data, other using in-domain monolingual data (Figure 2).

Automatic evaluation and result comparison with Google Translator is provided in the deliverable D5.4 (M29). Human evaluation comparing LetsMT! systems for business and financial news translation and Google Translator also was conducted; result analysis is provided in the deliverable D5.5 (M30).

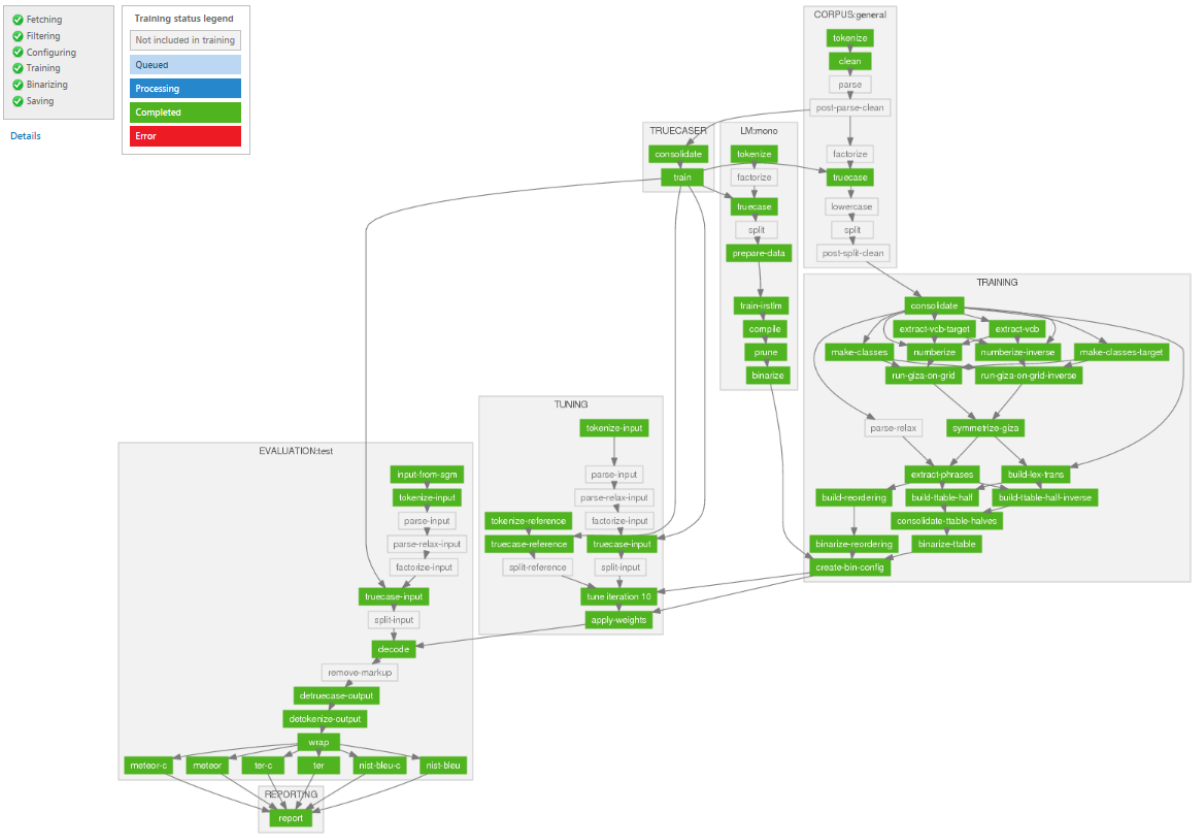

Figure 1. Training chart for SMT systems with one Language Model

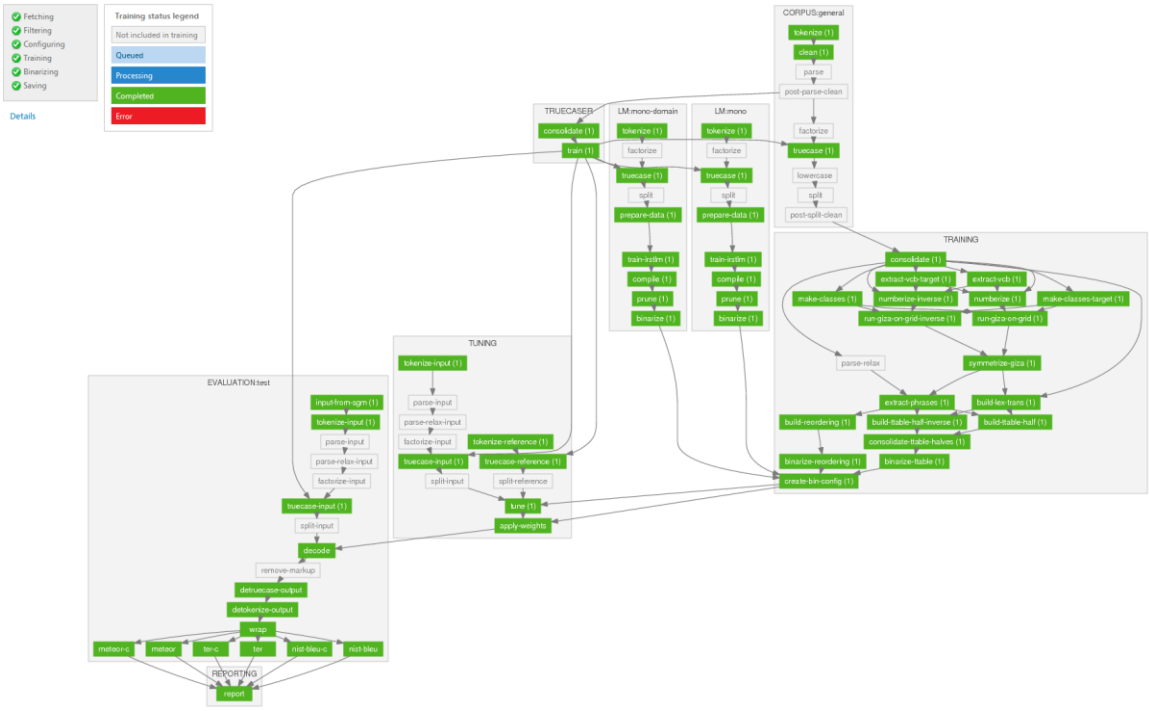

Figure 2. Training chart for SMT systems with two Language Models (general domain and in-domain)

In tables below we give detailed information about each trained SMT system – evaluation scores, parallel and monolingual corpora, tuning and evaluation data. Tables contain both (i) size of corpora used in training and (ii) size of actual data after filtering and removing duplicates.

Table 2. English-Danish system in business and financial domain

System:	English-Danish Finance IV
BLEU Score:	72.48
NIST Score:	10.7627
Monolingual data: (sentences)	267,875
Parallel data: (sentences)	112,836
Evolution set: (sentences)	794[3]
Tuning set: (sentences)	2,000[4]
Parallel corpora:	• Semlab Business News 1.1 (0.2 M)[5] • Semlab Business News 2 (0.2 M)[6]
Monolingual corpora:	• The target language part of parallel corpora • Finance (Danish monolingual) (0.23 M)[7]

[3] Evaluation, en-da finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-3c113c23-220a-4c9f-ab8b-68d920770b3b

[4] Automatically extracted from the training data

[5] https://www.letsmt.eu/CorporaDetails.aspx?id=c-f56fa25a-951c-47c0-a2a7-406adebd44dd

[6] https://www.letsmt.eu/CorporaDetails.aspx?id=c-6378b466-d3e3-4ffd-a938-e57824bfab17

[7] https://www.letsmt.eu/CorporaDetails.aspx?id=c-9e51c018-c94d-4afb-bf9c-79077199d45b

Table 3. English - Dutch system in business and financial domain

System:	English - Dutch Finance v3
BLEU Score:	62.98
NIST Score:	9.8749
Monolingual data: (sentences)	305,865
Parallel data: (sentences)	288,987
Evolution set: (sentences)	683[8]
Tuning set: (sentences)	1,631[9]
Parallel corpora:	• Semlab Business News 1.1 (0.3M)[10] • Semlab Business News 2 (0.3 M)[11] • Semlab Business News 3 (21.1K)[12]
Monolingual corpora:	• The target language part of parallel corpora

[8] Evaluation En-NL Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-7aa7d824-8dfa-4ed9-9e4f-98e0f1124d80

[9] Tuning En-NL Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-42d89e3b-0814-4abf-a392-76ceb8570743

[10] https://www.letsmt.eu/CorporaDetails.aspx?id=c-f56fa25a-951c-47c0-a2a7-406adebd44dd

[11] https://www.letsmt.eu/CorporaDetails.aspx?id=c-6378b466-d3e3-4ffd-a938-e57824bfab17

[12] https://www.letsmt.eu/CorporaDetails.aspx?id=c-57685a11-9a95-4026-81c8-568424f2bfaf

Table 4. English - Swedish system in business and financial domain

System:	English - Swedish Finance v3
BLEU Score:	65.47
NIST Score:	9.6325
Monolingual data: (sentences)	471,758
Parallel data: (sentences)	485,926
Evolution set: (sentences)	543[13]
Tuning set: (sentences)	2,000[14]
Parallel corpora:	• Semlab Business News 1.1 (0.5)[15] • Semlab Business News 2 (0.6 M)[16]
Monolingual corpora:	• The target language part of parallel corpora

[13] Evaluation En-Sv Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-45ffcfe8-c6a3-4ea5-89f2-56b68d8ecd17

[14] Automatically extracted from the training data

[15] https://www.letsmt.eu/CorporaDetails.aspx?id=c-f56fa25a-951c-47c0-a2a7-406adebd44dd

[16] https://www.letsmt.eu/CorporaDetails.aspx?id=c-6378b466-d3e3-4ffd-a938-e57824bfab17

Table 5. Dutch - English system in business and financial domain

System:	Dutch - English Finance v3
BLEU Score:	62.11
NIST Score:	9.9207
Monolingual data: (sentences)	1,087,088
Parallel data: (sentences)	288,283
Evolution set: (sentences)	683[17]
Tuning set: (sentences)	1,631[18]
Parallel corpora:	• Semlab Business News 1.1 (0.3M)[19] • Semlab Business News 2 (0.3 M)[20] • Semlab Business News 3 (21.1K)[21]
Monolingual corpora:	• The target language part of parallel corpora

[17] Evaluation En-NL Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-7aa7d824-8dfa-4ed9-9e4f-98e0f1124d80

[18] Tuning En-NL Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-42d89e3b-0814-4abf-a392-76ceb8570743

[19] https://www.letsmt.eu/CorporaDetails.aspx?id=c-f56fa25a-951c-47c0-a2a7-406adebd44dd

[20] https://www.letsmt.eu/CorporaDetails.aspx?id=c-6378b466-d3e3-4ffd-a938-e57824bfab17

[21] https://www.letsmt.eu/CorporaDetails.aspx?id=c-57685a11-9a95-4026-81c8-568424f2bfaf

Table 6. English - Czech system in business and financial domain

System:	English - Czech Finance v3
BLEU Score:	59.91
NIST Score:	9.1228
Monolingual data: (sentences)	2,115,677
Parallel data: (sentences)	2,290,644
Evolution set: (sentences)	558[22]
Tuning set: (sentences)	1,094[23]
Parallel corpora:	• DGT-TM-2007 2 [24] • DGT-TM-2011 [25] • Europarl v7 [26] • Semlab Business News 1.1 [27]
Monolingual corpora:	• LM-1: The target language part of general domain parallel corpora (DGT-TM-2007, DGT-TM-2011, Europarl) • LM-2: The target language part of in- domain parallel corpus: Semlab Business News 1.1

[22] Evaluation EN-CS Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-f3a0c6af-dba7-4aa0-80a3-a26a5b7620ac

[23] Tuning EN-CS Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-3620ecb3-f434-4978-9e67-2720d192c4bf

[24] https://www.letsmt.eu/CorporaDetails.aspx?id=c-dadc5f25-cb3d-4be6-9ce5-f8b381597837

[25] https://www.letsmt.eu/CorporaDetails.aspx?id=c-6b0bb47a-2684-4143-a11e-22619d228f08

[26] https://www.letsmt.eu/CorporaDetails.aspx?id=c-6475a45e-cddd-4601-96a4-0ea3449d5141

[27] https://www.letsmt.eu/CorporaDetails.aspx?id=c-f56fa25a-951c-47c0-a2a7-406adebd44dd

Table 7. English - Polish system in business and financial domain

System:	English - Polish Finance v3
BLEU Score:	53.05
NIST Score:	7.9928
Monolingual data: (sentences)	66,556
Parallel data: (sentences)	71,755
Evolution set: (sentences)	749[28]
Tuning set: (sentences)	1,192[29]
Parallel corpora:	• Semlab Business News 1.1 (0.5)[30] • Semlab Business News 3 (23.9K)[31]
Monolingual corpora:	• Target language part of parallel corpora

[28] LetsMT evaluation set: https://www.letsmt.eu/CorporaDetails.aspx?id=c-c490f017-43a6-4aba-a77b-1b2a6802e67d

[29] LetsMT development set: https://www.letsmt.eu/CorporaDetails.aspx?id=c-2d966e40-5845-4f1d-be9f-e141c39babb8

[30] https://www.letsmt.eu/CorporaDetails.aspx?id=c-f56fa25a-951c-47c0-a2a7-406adebd44dd

[31] https://www.letsmt.eu/CorporaDetails.aspx?id=c-57685a11-9a95-4026-81c8-568424f2bfaf

Table 8. English-Croatian system in business and financial domain

System:	English-Croatian Finance v3
BLEU Score:	17.42
NIST Score:	5.0307
Monolingual data: (sentences)	350,066
In-domain monolingual corpus (sentences):	415,806
Parallel data: (sentences)	360,579
Evolution set: (sentences)	394[32]
Tuning set: (sentences)	400[33]
Parallel corpora:	• Croatia Weekly News (1998-2000) (62.4K)[34] • Croatian-English Parallel Corpus (62.5K)[35] • EUBookShop (8K)[36] • hr-enWaC (91.1K)[37] • Semlab Business News 3 (12.5K)[38] • SETimes en-hr (0.2M)[39]
In-domain monolingual corpus	• Financial and business subcorpus of the hrWaC (0.4M)[40] • Semlab Business News 3 (12.5K)[41]
Monolingual corpora:	• Target language part of parallel corpora

[32] Evaluation En-Hr Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-fdc6f728-b36e-4a34-a009-fbad838222ec

[33] Tuning En-Hr Finance, validated: https://www.letsmt.eu/CorporaDetails.aspx?id=c-3690ffa3-eab6-4a8d-811b-63028e0efa03

[34] https://www.letsmt.eu/CorporaDetails.aspx?id=c-a66baa0c-76c9-4b58-8d84-f7d1bebea624

[35] https://www.letsmt.eu/CorporaDetails.aspx?id=c-405e54ec-1abc-4cf9-98a7-15894eeec229

[36] https://www.letsmt.eu/CorporaDetails.aspx?id=c-b342978f-8e85-4705-86a2-63cee2a34956

[37] https://www.letsmt.eu/CorporaDetails.aspx?id=c-98823ac6-fba6-48db-a916-6f94acd212c8

[38] https://www.letsmt.eu/CorporaDetails.aspx?id=c-57685a11-9a95-4026-81c8-568424f2bfaf

[39] https://www.letsmt.eu/CorporaDetails.aspx?id=c-a861008f-aff9-4977-8454-437cba41b6c8

[40] https://www.letsmt.eu/CorporaDetails.aspx?id=c-bf8df7eb-9ffd-411f-a046-0a8b9428aa04

[41] https://www.letsmt.eu/CorporaDetails.aspx?id=c-57685a11-9a95-4026-81c8-568424f2bfaf

3 Comparison with Google Translate

A benchmark study of LetsMT! business and financial news translation systems against Google Translate has been done. This section gives just a short summary of this comparison with Google Translate. Detailed results are presented in deliverables:

- D5.4 Automatic evaluation report of business and financial news SMT
- D5.5 Evaluation report on usability of SMT in business and financial news translation

According to the automatic evaluation measures all the LetsMT! systems except for English-Croatian get a significantly better score than the output translation by *Google Translate.*

The results according to the respective quality of the machine translations and the event detection yield the following graphs:

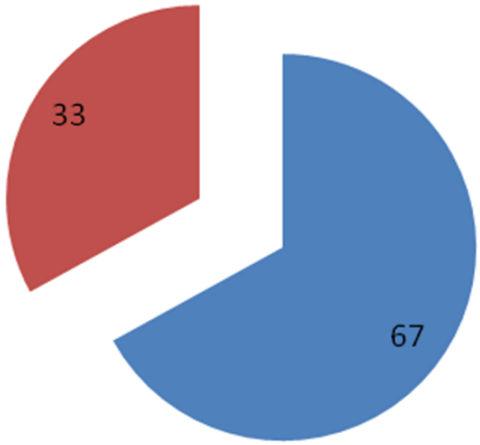

Figure 3. Percentage 33% wrong 67% good of correct ViewerPro events with Google Translate over the dataset of 500 random chosen messages**.**

The findings indicate a 67% adequate event translation with Google Translate against 80% adequate event translation with LetsMT.

4 Conclusions

All SMT systems necessary for evaluation of the LetsMT! platform in business and financial news translation scenario are trained. Trained SMT systems are automatically evaluated in Task 5.4/D5.4 (M29) and usability of SMT in business and financial news translation are assessed in Task 5.5/D5.5 (M30).