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. 1.0

30/06/2011

Document Information

Deliverable number:	D5.3
Deliverable title:	SMT systems trained for business and financial news translation - draft
Due date of deliverable according to DoW:	30/06/2011
Actual submission date of deliverable:	30/06/2011
Main Author(s):	SemLab: Dohmen
Participants:	-
Reviewer	Tilde
Workpackage:	WP5
Workpackage title:	MT usage in news translation: facilities and evaluation
Workpackage leader:	SEM
Dissemination Level:	PU
Version:	V1.0
Keywords:	SMT training, Finance, Business

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval Level
0.1	15/06/2011	Draft	SEM	-	Ready for comments
0.2	28.06.2011	Draft	Tilde	Document formatting, explanation of NIST/BLEU scores.	Ready for comments
1.0	30.06.2011	Final	Tilde	-	Final document version

Table of Contents

1 Business and Financial Corpora

The initial corpora that were trained for use in the LetsMT! widget, are in the business and finance domain. This topic is chosen because the widget will be showcased and tested on the financial news website www.newssentiment.eu (which is a website that gathers the sentiment in financial news for all major European equities and from over 53 live news sources a.o. Dow Jones news stream) and also in following websites: http://www.thelocal.se/news/0/, http://www.dutchnews.nl/, http://www.praguepost.com/business/, http://www.thenews.pl/, http://blogs.wsj.com/new-europe/.

The training of the six small European languages (Polish, Danish, Swedish, Dutch, Croatian and Czech) was done using the LetsMT! system by Tilde, based on the initially available corpora (gathered in WP4, see deliverable D4.5).

The size of the corpora were initially estimated on the frequency per language in financial & business news: Dutch (at least 2.3 million running words), Swedish (at least 1 million running words), Polish (at least 8.7 million running words), Danish (at least 0.7 million running words), Czech (at least 0.6 million running words), Croatian (unknown).

The Description of Work (DOW) of LetsMT! erroneously mentions Slovakian instead of the intended Swedish. In the subsequent sections Swedish is mentioned, as originally intended. Since none of the partners have knowledge of Slovakian and since there is little demand for business news information in Slovakian, we have decided to stick with the originally planned Swedish corpus.

2 Training the LetsMT! system

In the sections below the training status and additional information of each financial language corpus is given. The green dot indicates that the system is running (red would mean that it is still training, yellow that the system is idle).

2.1 *BLEU & NIST scores*

The BLEU (Papineni et al. 2002) and NIST (Doddington 2002) scores are indications of the system's translation quality.

From Wikipedia: *BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text that has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human. Scores are calculated for individual translated segments by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical correctness are not taken into account. BLEU's output is always a number between 0 and 1.*

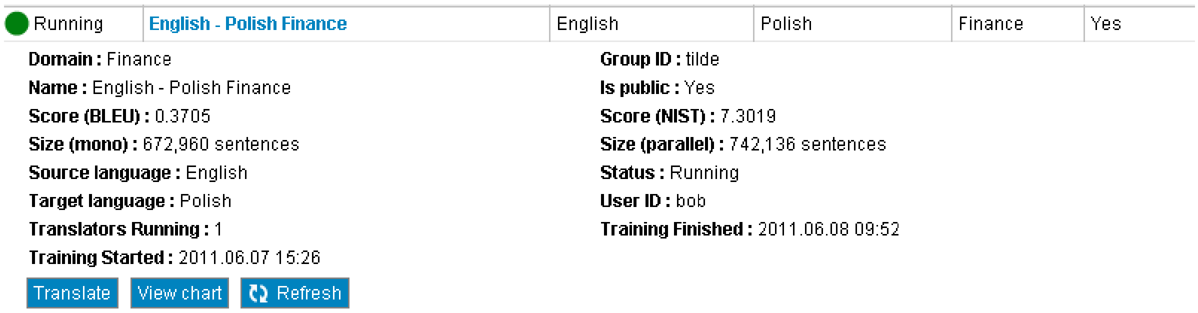
The range for BLEU results is from 0 – 1. The score indicates how similar the translation and the reference text is. Values closer to 1 represent more similar texts.

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.

From Wikipedia: *NIST (National Institute of Standards and Technology) It is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is. That is to say when a correct n-gram is found, the rarer that n-gram is, the more weight it will be given. For example, if the bigram "on the" is correctly matched, it will receive lower weight than the correct matching of bigram "interesting calculations", as this is less likely to occur. NIST also differs from BLEU in its calculation of the brevity penalty insofar as small variations in translation length do not impact the overall score as much.*

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.

2.2 Polish-English



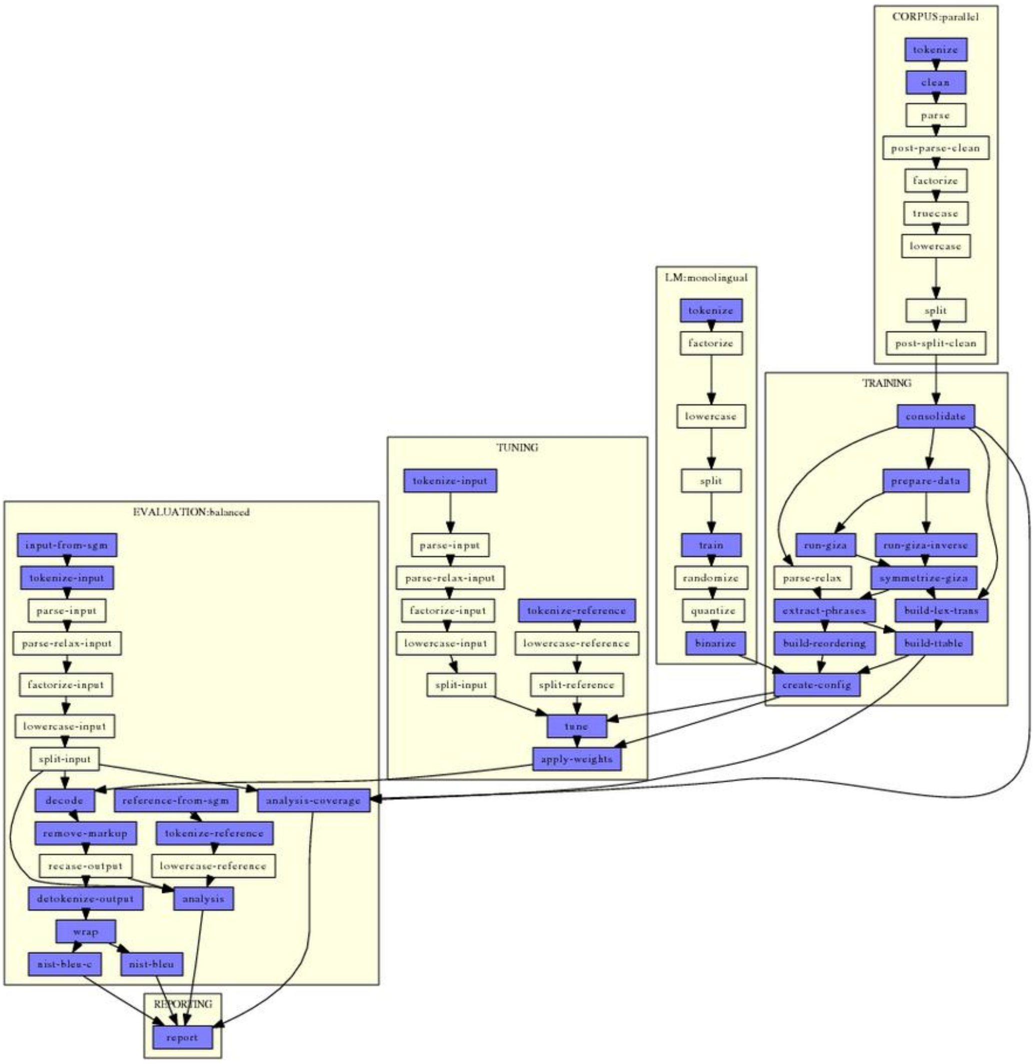
● Running	English - Polish Finance	English	Polish	Finance	Yes

Domain : Finance
Name : English - Polish Finance
Score (BLEU) : 0.3705
Size (mono) : 672,960 sentences
Source language : English
Target language : Polish
Translators Running : 1
Training Started : 2011.06.07 15:26

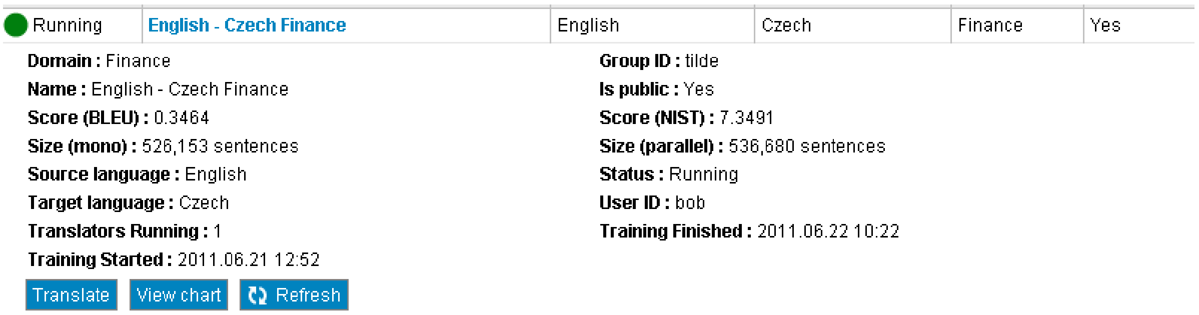
Group ID : tilde
Is public : Yes
Score (NIST) : 7.3019
Size (parallel) : 742,136 sentences
Status : Running
User ID : bob
Training Finished : 2011.06.08 09:52

The screenshot above is from the LetsMT! web interface (this can be found on https://demo.letsmt.eu) and shows the status of the system's training, including the corpus size and BLEU/NIST scores that indicate the system's translation quality. In this case, Polish-English Finance training has been completed and the MT-system is currently running.

The image below shows the training phases and individual training steps that the Polish-English Finance training has undergone. This chart looks the same for every trained language pair and is therefore only shown once.



2.3 *Czech-English*



● Running	English - Czech Finance	English	Czech	Finance	Yes

Domain : Finance

Name : English - Czech Finance

Score (BLEU) : 0.3464

Size (mono) : 526,153 sentences

Source language : English

Target language : Czech

Translators Running : 1

Training Started : 2011.06.21 12:52

Group ID : tilde

Is public : Yes

Score (NIST) : 7.3491

Size (parallel) : 536,680 sentences

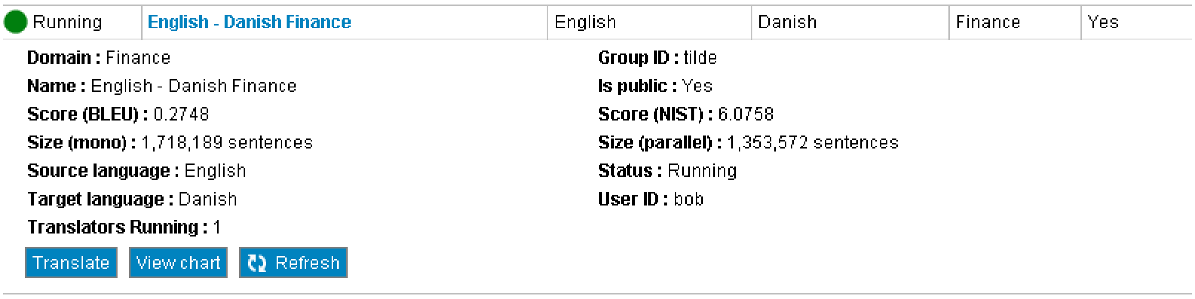
Status : Running

User ID : bob

Training Finished : 2011.06.22 10:22

Czech-English Finance training has been completed and the MT-system is currently running.

2.4 Danish-English



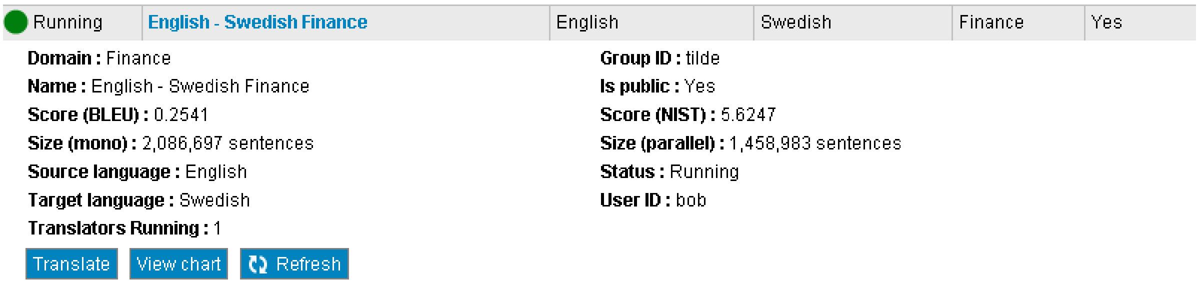
● Running	English - Danish Finance	English	Danish	Finance	Yes

Domain : Finance
Name : English - Danish Finance
Score (BLEU) : 0.2748
Size (mono) : 1,718,189 sentences
Source language : English
Target language : Danish
Translators Running : 1

Group ID : tilde
Is public : Yes
Score (NIST) : 6.0758
Size (parallel) : 1,353,572 sentences
Status : Running
User ID : bob

Danish-English Finance training has been completed and the MT-system is currently running.

2.5 Swedish-English



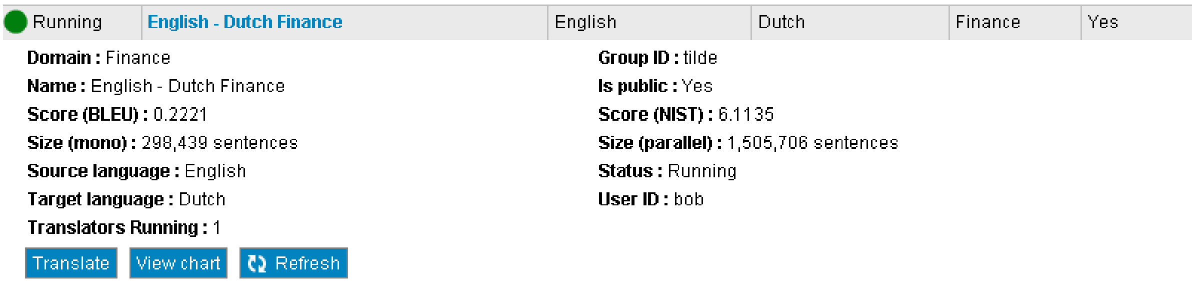
● Running	English - Swedish Finance	English	Swedish	Finance	Yes

Domain : Finance
Name : English - Swedish Finance
Score (BLEU) : 0.2541
Size (mono) : 2,086,697 sentences
Source language : English
Target language : Swedish
Translators Running : 1

Group ID : tilde
Is public : Yes
Score (NIST) : 5.6247
Size (parallel) : 1,458,983 sentences
Status : Running
User ID : bob

Swedish-English Finance training has been completed and the MT-system is currently running.

2.6 Dutch-English



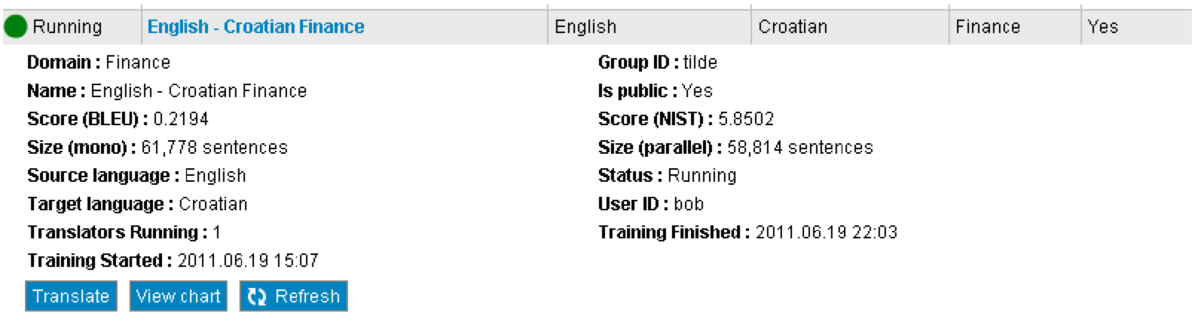
● Running	English - Dutch Finance	English	Dutch	Finance	Yes

Domain : Finance
Name : English - Dutch Finance
Score (BLEU) : 0.2221
Size (mono) : 298,439 sentences
Source language : English
Target language : Dutch
Translators Running : 1

Group ID : tilde
Is public : Yes
Score (NIST) : 6.1135
Size (parallel) : 1,505,706 sentences
Status : Running
User ID : bob

Dutch-English Finance training has been completed and the MT-system is currently running.

2.7 *Croatian-English*



● Running	English - Croatian Finance	English	Croatian	Finance	Yes

Domain : Finance
Name : English - Croatian Finance
Score (BLEU) : 0.2194
Size (mono) : 61,778 sentences
Source language : English
Target language : Croatian
Translators Running : 1
Training Started : 2011.06.19 15:07

Group ID : tilde
Is public : Yes
Score (NIST) : 5.8502
Size (parallel) : 58,814 sentences
Status : Running
User ID : bob
Training Finished : 2011.06.19 22:03

Croatian-English Finance training has been completed and the MT-system is currently running.

3 Future Developments

The size of all corpora will be extended over the next 6 months with an additional body of parallel texts. Since MT training quality depends on the quality and size of the parallel corpora, this is the main goal.

4 Conclusion

The collection and training of the parallel business and finance coprora has been going according to plan and on schedule. This enables the integration of the translation widget to proceed as planned.

5 References

K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: a method for automatic evaluation of machine translation, in Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL), 2002

G. Doddington, Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in Proceedings of HLT-02, 2002.