This document is part of the Project "Machine Translation Enhanced Computer Assisted Translation (MateCat)", funded by the 7th Framework Programme of the European Commission through grant agreement no.: 287688.

Machine Translation Enhanced

Computer Assisted Translation

D5.6 - Evaluation Benchmarks

Authors: Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Alessandro Cattelan

Dissemination Level: Public

Date: November 9th, 2014

Grant agreement no.	287688
Project acronym	MateCat
Project full title	Machine Translation Enhanced Computer Assisted Translation
Funding scheme	Collaborative project
Coordinator	Marcello Federico (FBK)
Start date, duration	November 1st 2011, 36 months
Dissemination level	Public
Contractual date of delivery	October 31st, 2014
Actual date of delivery	November 9th, 2014
Deliverable number	D5.6
Deliverable title	Evaluation Benchmarks
Type	Report
Status and version	Final
Number of pages	10
Contributing partners	FBK, TRANSLATED
WP leader	TRANSLATED
Task leader	TRANSLATED
Authors	Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Alessandro Cattelan
Reviewers	Manuela Speranza
EC project officer	Alexandra Wesolowska
The partners in MateCat are:	Fondazione Bruno Kessler (FBK), Italy
	Université Le Mans (LE MANS), France
	The University of Edinburgh (UEDIN)
	Translated S.r.l. (TRANSLATED)

For copies of reports, updates on project activities and other MateCat-related information, contact:

FBK
MateCat
Manuela Speranza manspera@fbk.eu
Povo - Via Sommarive 18 Phone: +39 0461 314 521
I-38123 Trento, Italy Fax: +39 0461 314 591

Copies of reports and other material can also be accessed via http://www.matecat.com

Executive Summary

This document reports on the benchmarks publicly released by the project. They are:

- `Field-test data`: documents without copyright issues employed in the field-tests; the benchmark includes the source text, the human reference translation (if any), the suggestion chosen by the translator to post-edit and the final post-edition
- `BinQE`: collection of source/automatic translations in different language pairs with binary labels (good/bad), developed for quality estimation tasks
- `BitterCorpus`: collection of parallel English-Italian documents in the IT domain, with domain-specific terms manually marked and aligned
- `Word-alignment Gold Reference`: collection of human-checked word-alignment of English-Italian sentence pairs in the Legal domain

Table of Contents

1 Introduction 5
2 Field-test Data 5
3 BinQe 7
4 BitterCorpus 7
5 Word-alignment Gold Reference 8
References 9

1 Introduction

The goal of the dissemination and exploitation activities of the MateCat project is to promote its outcomes among the scientific, industrial and user communities. It is achieved through the publication of technical papers, the presentation of the MateCat tool versions and the field test results among the industrial players, and the promotion of the MateCat tool among professional translators as well as occasional translators of Web communities. The implemented software is also documented and distributed as open source. Finally, in order to allow for an effective exploitation and deployment of the project results, a set of benchmarks is made publicly available, consisting of LRs prepared and collected for accomplishing the activities of the project.

This deliverable describes such benchmarks, namely the Field-test data, BinQE, BitterCorpus and the Word-alignment Gold Reference.

2 Field-test Data

The MateCat field tests were conducted on the IT and LEGAL domains; TED[1] domain was added in the last experiments, following a recommendation raised from the project review related to the period M13 to M24. IT documents were selected from the collection of real translation projects commissioned to Translated, the industrial partner of the project; therefore, they are proprietary of customers and cannot be released. On the other hand, all LEGAL and TED documents translated during the field tests can be freely distributed, so they have been packaged to form this benchmark.

Typically, field tests were organized over two days in which different documents had to be translated by four professional translators. During the first day, translators received automatically generated translations of segments of the first document (doc1) by either a reference MT engine or the TM; during the second day, suggestions related to the second document (doc2) came from either an enhanced MT engine or again from the TM. The benchmark is built on the documents employed in the five field tests grouped as shown in the rows of Table 1.

For each field test, domain and language pair, Table 1 provides the number of segments and tokens of the English side of the documents translated during the two days. Note that the LEGAL document involved in the four groups 02 to 05 is the same and shared among the two language-pairs. The small differences between doc1 texts is due to a few segments being

[1] www.ted.com

unavailable for all languages. doc2 of groups 04 and 05 is a subset of that of group 02. The field test corresponding to group 03 lasted only one day.

As already stated, the number of translators who post-edited the documents is four; there is one exception which regards the doc2 of the LEGAL English-{Italian,French} tasks: in these cases, the work of only three translators is worthy to be released.

Field Test	LEGAL								TED			
	English-Italian				English-French				English-French			
	doc1		doc2		doc1		doc2		doc1		doc2	
	seg	wrds	seg	wrds	seg	wrds	seg	wrds	seg	wrds	seg	wrds
group01	91	2955	90	3004	91	2955	90	3004	200	3322	165	2967
group02	133	3084	472	10820	134	3086	472	10820	200	3322	165	2967
group03	133	3084			134	3086						
group04	132	3105	152	3633	132	3113	152	3633				
group05	132	3105	152	3633	132	3113	152	3633				

Table 1: Statistics on the field-test data sets.

The LEGAL document[2] shared by the four groups 02 to 05 is taken from the European Union law[3], for which translations into the four languages of interest for MateCat (French, German, Italian and Spanish) were available. The document was pre-processed so that the segments of the four versions were all aligned. The full document consists of about 600 segments and 13,900 English words, and was split into two portions: one for the first day session (doc1: 132-134 segments) and the rest used somehow in the second session (doc2). More details are provided in the three project reports on the lab and field tests (deliverables D5.3, D5.4 and D5.5).

TED documents were taken from the WIT3 release[4] prepared for the IWSLT 2014[5] evaluation campaign; in particular, the texts translated in the two sessions - and shared among groups 01 and 02 - are the transcriptions of two talks (numbers 1181 and 1427) taken from development sets.

For each entry of the table, the benchmark releases the set of documents involved in the job done by each of the four (three) translators; in particular, for translator N (N=1,2,...), they are:

[2] 2013/488/EU: "Council Decision of 23 September 2013 on the security rules for protecting EU classified information".
[3] eur-lex.europa.eu
[4] wit3.fbk.eu
[5] workshop2014.iwslt.org

`TON.en`	source text
`TON.fr|it`	human reference (when available)
`TON.fr|it.sugg`	suggestion chosen by the translator to post-edit
`TON.fr|it.pe`	post-edition

The archive `MATECAT-FieldTestsBenchmark.tgz`, including the field test documents and a more detailed description, is available for download at:

http://www.mt4cat.org/benchmarks/matecat-post-edits

3 BinQe

Machine Translation (MT) Quality Estimation (QE) (Specia et al., 2009, Specia et al., 2010, Mehdad et al., 2012, Turchi et al., 2014) is the task of determining the quality of an automatic translation given its source sentence and without recourse to reference translations. While most of the currently available datasets are obtained through manual annotation of (source,target) sentence pairs with continuous scores or Likert values (e.g. wrt a 5-point scale where 1="Incomprehensible" and 5="Flawless translation), little has been done to produce binary datasets with "good" (useful, or suitable for post-editing) vs "bad" (useless, needs complete rewriting) judgements. This kind of judgements is particularly useful to train QE models for specific applications such as the integration in a Computer-assisted translation environment where a sharp distinction between “good” and “bad” translation suggestions is needed.

BinQE (Turchi and Negri 2014) is a collection of binary QE datasets for different language pairs, where the labels have been automatically produced by applying the method described in (Turchi et al., 2013). More specifically, BinQE contains:

- 2,754 English-Spanish news sentences from the WMT 2013 datasets
- 10,881 French-English news sentences from the corpus described in (Potet et al. 2010)
- 1,261 English-Italian sentences from the legal domain collected within MateCat (Federico et al., 2014)

BinQE is available for download at:

http://www.mt4cat.org/benchmarks/binqe

4 BitterCorpus

BitterCorpus (Arcan et al., 2014) is a collection of parallel English-Italian documents in the IT domain where domain-specific terms have been manually marked and aligned. The documents are extracted from the GNOME and the KDE data collections. They contain 874 domain-specific bilingual terms in total. More specifically, BitterCorpus contains:

- GNOME Corpus
  It contains 55 parallel documents extracted from the Gnome manual documentation (IT domain). Three annotators, fluent in English and Italian, have been selected to annotate the documents with domain-specific terms. In total, they have annotated 313 Italian and 282 English terms and 237 bilingual domain-specific terms.

- KDE Corpus
  It contains one parallel document extracted from the KDE manual documentation (IT domain), whereby the document is made of 100 lines of text. Three annotators, fluent in English and Italian, have been selected to annotate the documents with domain-specific terms. In total, they have annotated 628 Italian and 628 English terms, and 637 bilingual domain-specific terms.

BitterCorpus is available for download at:

http://www.mt4cat.org/benchmarks/bittercorpus

5 Word-alignment Gold Reference

In addition to the enhanced word-aligner we proposed in (Farajian, 2014) and described in the MateCat deliverable D6.5 Open Source Distribution, we developed a gold reference for the word-alignment task. This benchmark contains 200 English-Italian sentence pairs in the LEGAL domain extracted from the JRC-Acquis Corpus, and their word-to-word alignments produced by two professional translators.

A script for evaluating the alignment is also included.

The archive `MATECAT_WordAlignmentGoldReferenceBenchmark.tgz` containing the benchmark is available for download here:

http://www.mt4cat.org/benchmarks/word-alignment-gold-reference

References

Arcan, Mihael, Marco Turchi, Sara Tonelli, Paul Buitelaar. 2014. “Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment”. Proceedings of AMTA, 2014.

Farajian, M. Amin, Nicola Bertoldi and Marcello Federico. 2014. “Online Word Alignment for Online Adaptive Machine Translation”. Proceedings of the Workshop on Humans and Computer-assisted Translation, co-located with the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden.

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. “THE MATECAT TOOL.” Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin, Ireland.

Mehdad, Yashar, Matteo Negri, and Marcello Federico. 2012. “Match without a Referee: Evaluating MT Adequacy without Reference Translations.” Proceedings of the 7th Workshop on Statistical Machine Translation (WMT2012). Montreal, Canada.

Potet Marion, Emmanuelle Esperança-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. “Collection of a Large Database of French-English SMT Output Corrections”. Proceedings of LREC 2012.

Specia, Lucia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. “Estimating the sentence-level quality of machine translation systems.” Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT’09). Barcelona, Spain.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010. “Machine Translation Evaluation versus Quality Estimation.” Machine Translation, 24(1).

Turchi, Marco, Matteo Negri, and Marcello Federico. 2013. “Coping with the Subjectivity of Human Judgements in MT Quality Estimation.” Proceedings of the 8th Workshop on Statistical Machine Translation (WMT’13). Sofia, Bulgaria.

Turchi, Marco, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. "Adaptive Quality Estimation for Machine Translation". Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland.

Turchi, Marco and Matteo Negri. 2014. "Automatic Annotation of Machine Translation Datasets with Binary Quality Judgements." Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.