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

D6.5 - Open source distribution

Authors: Nicola Bertoldi, Marco Turchi, Ulrich Germann, Frederic Blain, Holger Schwenk, Anthony Rousseau, 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	D6.5
Deliverable title	Open Source Distribution
Type	Other
Status and version	Final
Number of pages	12
Contributing partners	FBK, TRANSLATED, LIUM, UEDIN
WP leader	FBK
Task leader	FBK
Authors	Nicola Bertoldi, Marco Turchi, Ulrich Germann, Frederic Blain, Holger Schwenk, Anthony Rousseau, 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 software created by the MateCat Consortium during the whole duration of the project and publicly released. The delivered software packages are:

- MateCat Tool
- Adaptive MT Server
- Online-adaptive Moses with cache-based models
- Online-adaptive Moses with suffix-array models
- OnlineMGIZA++
- Bitext-tokaligner
- AQET
- CSLM Toolkit
- XenC
- MT-EQuAl
- BATMAN

All software packages are distributed in open source, each of them with its own license governing its use and redistribution. The policy followed was to release all the software that has been completely developed in MateCat under an LGPL license while maintaining the original open source license for software that has been developed from existing open source software.

Table of Contents

1 Introduction

The MateCat project addressed the integration of machine translation (MT) and human translation within the computer aided translation (CAT) framework. The ultimate goal was to improve productivity of professional translators and to enhance their work experience with MT. To achieve this goal and advance state-of-the-art, the project focused on making MT technology aware of how it is used, attuning itself to the domain, adapting to the corrections and implicit feedback of the translators, and providing them with useful information. Pursuing this objective was and still is definitely relevant for everyone who works in the translation field, both in industry and in research, as well as for occasional translators.

Hence, the ultimate goal of the dissemination and exploitation activities of the MateCat project was to promote its outcomes among the scientific, industrial, and user communities. This goal has been pursued through the publication of technical papers in the top scientific venues and journals, 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. Finally, all the implemented software has been documented and distributed as open source, to foster their rapid exploitation.

This deliverable enumerates all toolkits developed along the whole duration of the project, providing for each of them a brief description, pointers to full installation, configuration and usage documentation, and the link where to download the product.

2 Open Source Software

2.1 MateCat Tool

MateCat is an enterprise-level, web-based CAT tool (Federico et al., 2014) designed to make post-editing and outsourcing easy and to provide a complete set of features to manage and monitor translation projects.

Thanks to the integration of the largest collaborative translation memory and statistical machine translation, users will likely get significantly more matches than with other CAT tools and translate faster.

MateCat is a free, open source software released under the LGPL license.

Source code is available here:

https://www.assembla.com/code/matecat_source/git/nodes

All information needed to install and configure MateCat is available here:

http://www.matecat.com/installation-guide

The MateCat Tool is distributed under the GNU Lesser General Public License (LGPL).

2.2 Adaptive MT server

Adaptive MT server is a translation web service able to return a list of translation alternatives for a given input, automatically generated by an MT engine, and to adapt on-the-fly the MT engine models according to any provided feedback, consisting of a parallel sentence pair, i.e. a source text and its translation. Adaptive MT server is also able to support terminology, either pre-loaded or inserted on the-fly.
Currently, adaptive MT server is connected with the well-known state-of-the-art phrase-based SMT Moses toolkit, supporting the online adaptation via cache-based models. In order to perform the online adaptation of the models, adaptive MT server also requires a word alignment software. Currently, it supports both the online MGIZA++ (see below) and Bitext-tokaligner (see below).
Furthermore, adaptive MT server interfaces with AQET (see below), to compute the quality estimation of the automatic translations, and to adapt these values score to the user feedback.

Source code and all instructions needed to install and configure the adaptive MT server are available here:

http://www.mt4cat.org/software/mt-server

Adaptive MT server is distributed under the GNU Lesser General Public License (LGPL).

2.3 Online-adaptive Moses with cache-based models

Online-adaptive Moses (Bertoldi, 2014) is an enhanced version of the well-known state-of-the-art SMT Moses[1] Toolkit, which permits the dynamic (on-the-fly) adaptation of its statistical models without the need of their reloading, according to any suggestions coming from users.
Online adaptation is achieved by means of cache-based language and translation models, which reward their content using either a parameterizable time-decaying scoring function or pre-defined constant values. The caches can be populated at any time from input using an xml-based annotation or pre-populated from file.
Online-adaptive Moses has been successfully integrated in adaptive MT server connected to the MateCat Tool; in this way, it adapts to user post-editing, and supports the usage of terminology.

Source code is available in the branch “dynamic-models” of the Github repository for Moses:

[1] http://www.statmt.org/moses

https://github.com/moses-smt/mosesdecoder/tree/dynamic-models

Details about configuration and usage of the cache-based translation and language models are available at these two URLs, respectively:

http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc14
http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc15

Moses is distributed under the GNU Lesser General Public License (LGPL).

2.4 Online-adaptive Moses with suffix-array models

The phrase-based version of Moses has been also enhanced with a fast and reliable implementation of virtual phrase tables based on sampling word-aligned bitexts at query time (Germann, 2014)[2]. Following the approach suggested by Callison-Burch et al. (2005), the training data is indexed for full text search via suffix arrays; phrase table entries are constructed on-the-fly by extracting translations from a sufficiently large sample of source phrase occurrences in the word-aligned parallel data. This approach precludes computation of certain feature values included in conventional phrase tables, such as smoothed conditional translation probabilities; for this reason alternative feature functions were developed in the course of this work, which ensure the same level of translation quality that can be achieved with conventional phrase tables. In terms of lookup speed, the new implementation outperforms existing conventional implementations of phrase tables.
The underlying word-aligned parallel corpus can be amended dynamically with low latency, for example by feeding back translation output after it has been post-edited by translation professionals. New additions to the corpus can be exploited for future translations immediately.

Source code is available in the master branch of the Github repository for Moses:

https://github.com/moses-smt/mosesdecoder

Usage information about the suffix-array-based phrase table is available here:

http://www.statmt.org/moses/?n=Moses.PhraseDictionaryBitextSampling

Moses is distributed under the GNU Lesser General Public License (LGPL).

2.5 OnlineMGIZA++

OnlineMGIZA++ is an enhanced version of MGIZA++ (Gao et al., 2008) that is well-suited for the real-time applications in which the sentence pairs are required to be word-aligned,

[2] As a fundamental contribution to Moses as a back-end MT server in post-editing frameworks, this work spanned several projects funded at the University of Edinburgh under FP7: MateCat, CasMaCat (Grant Agreement No. 287576), and ACCEPT (Grant Agreement No. 288769).

one at a time. Nevertheless, it inherited the functionality to train alignment models from a parallel training corpus.
According to a set of pre-trained alignment models, onlineMGIZA++ word-aligns any new sentence pair input from stdin, possibly discovering links for unknown words. This enhanced version of MGIZA++ shows better performance than other well-known word-alignment toolkits (Farajian et al., 2014).

Source code and instructions to install and configure onlineMGIZA++ are available here:

http://www.mt4cat.org/software/onlineMGIZA++

As a derivative work of MGIZA++, onlineMGIZA++ is distributed under the GNU General Public License version 2.0 (GPLv2).

2.6 Bitext-tokaligner

Bitext-tokaligner (Blain, 2012) is a pivot-based word aligner; this software consists of a Perl script that implements a simple algorithm to align at word level a source sentence and its human-generated reference using a raw translation version as pivot. It has been designed to align sentence pairs in a real-time post-editing context. Both its usability and its limited time consumption make this software a suitable approach for this purpose.
The script relies on the Moses toolkit, Moses models including word-to-word alignments, and TERcpp, a c++ implementation of the TER.

Source code and usage instructions are available here:

https://github.com/fredblain/bitext-tokaligner

Bitext-tokaligner is distributed under the GNU Lesser General Public License (LGPL).

2.7 AQET

AQET (Adaptive Quality Estimation Tool) (Turchi et al., 2014) is an open-source package for performing Quality Estimation for Machine Translation, i.e. the task of determining the quality of an automatic translation given its source sentence and without recourse to reference translations (Specia et al. 2009; Specia et al., 2010; Mehdad et al., 2012). AQET is able to continuously learn from post-edited sentences and is reactive and robust to user and domain changes. AQET has been developed to support professional translators during their daily work and it is suitable for being embedded in a CAT tool. The current version (v1.0) supports three online machine learning algorithms: Online Support Vector Regression, Online Gaussian Process and Passive-Aggressive.

A detailed description of the software, including installation and configuration instructions, together with the source code is available here:

http://www.mt4cat.org/software/aqet

AQET is distributed under the GNU General Public License version 3 (GPLv3).

2.8 CSLM Toolkit

During the project, the existing CSLM toolkit (Schwenk, 2010; Schwenk 2013) has been integrated in Moses, extended with project adaptation functions and extensively experimented. CSLM is an open-source software which implements the so-called continuous space language and translation model. The basic idea of this approach is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown events can be expected. A neural network can be used to simultaneously learn the projection of the words onto the continuous space and to estimate the n-gram probabilities. This is still a n-gram approach, but the LM probabilities are interpolated for any possible context instead of backing-off to shorter contexts. This approach was successfully used in large vocabulary continuous speech recognition and in phrase-based SMT systems.

There is a strongly increasing interest in the use of neural networks for SMT and other NLP applications. As far as we know, the CSLM toolkit is the only freely available tool which allows to train neural network language model on large corpora in an efficient way. It supports training of the models on standard machines and GPU cards.

The source code, together with a tutorial which outlines the training of neural network language models and the procedure to use it with Moses, is freely available at the following location:

http://www-lium.univ-lemans.fr/cslm/

The CSLM toolkit is distributed under the GNU General Public License version 3 (GPLv3).

2.9 XenC

XenC (Rousseau, 2013) is a C++ pre-processing tool aimed at selecting textual data. It has applications to NLP and particularly Statistical Machine Translation (SMT) and Automatic Speech Recognition (ASR). It can perform language-independent monolingual as well as bilingual data selection. The goal of XenC is to allow selection of relevant data regarding a given task or subject, which then will be used to build the statistical models for a NLP system.

It implements reputed methods from (Moore, 2010) and (Axelrod et al., 2011) based on cross-entropy differences between in-domain data and noisy out-of-domain data.

The source code and the detailed description, including the instructions for installation, configuration and usage are available here:

https://github.com/rousseau-lium/XenC

XenC is distributed under the GNU General Public License version 3 (GPLv3).

2.10 MT-EQuAl

MT-EQuAl (Girardi et al., 2014) is a toolkit for the human assessment of Machine Translation output.
The web-based toolkit provides an annotation interface to carry out three annotation tasks, namely quality rating of translations (e.g. adequacy/fluency, relative ranking), annotation of translation errors, and word-alignment.
MT-EQuAl supports the following browsers: Chrome, Safari, Firefox, and Internet Explorer.

The source code and the detailed description, including the instructions for installation, configuration and usage and the annotation guidelines, are available here:

http://www.mt4cat.org/software/mt-equal

MT-EQuAl is distributed under the Apache License, Version 2.0.

2.11 BATMAN

BilinguAl TerM AligNer (BATMAN) (Arcan et al, 2014) is an open-source tool for aligning monolingual terminology, extracted from parallel texts, across different languages. BATMAN requires in input monolingual terms from the source and target languages and the parallel documents from where the terms have been extracted. As a result, it provides a list of aligned bilingual terminology.

The source code and the usage instructions are available here:

http://www.mt4cat.org/software/batman

BATMAN is distributed under the GNU Lesser General Public License (LGPL).

References

Arcan, Mihael, Marco Turchi, Sara Tonelli and Paul Buitelaar. 2014. “Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment”. Proceedings of the Association for Machine Translation in the Americas (AMTA ‘14), Vancouver, Canada, pp 54-68.

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. “Domain adaptation via pseudo in-domain data selection”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 355–362.

Bertoldi, 2014, “Dynamic Models in Moses for Online Adaptation”. The Prague Bulletin of Mathematical Linguistics, 101, pp. 7–28.

Blain, Frederic, Holger Schwenk, and Jean Senellart. 2012. "Incremental Adaptation Using Translation Information and Post-Editing Analysis", Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Hong-Kong, China, pp. 234-241.

Callison-Burch, Chris, Colin Bannard, and Josh Schroeder. 2005. “Scaling Phrase-based Statistical Machine Translation to Larger Corpora and Longer Phrases”. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL ’05), Ann Arbor, Michigan, USA, pp. 255–262.

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, pp. 84–92.

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, pp. 129–132.

Germann, Ulrich. 2014. "Dynamic Phrase Tables for Statistical Machine Translation in an Interactive Post-editing Scenario". Proceedings of the AMTA 2014 Workshop on Interactive and Adaptive Machine Translation. Vancouver, BC, Canada, pp. 20–31.

Girardi, Christian, Luisa Bentivogli, Mohammad Amin Farajian and Marcello Federico. 2014, "MT-EQuAl: a Toolkit for Manual Assessment of Machine Translation Output". Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin, Ireland, pp. 120–123.

Moore, Robert C. and William Lewis. 2010. "Intelligent selection of language model training data". In Proceedings of the ACL Conference Short Papers, Uppsala, Sweden, pp. 220–224.

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, pp. 171–180.

Qin, Gao and Stephan Vogel. 2008. "Parallel Implementations of Word Alignment Tool". Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing. Columbus, US-OH, pp. 49–57.

Rousseau, Anthony. 2013. "XenC: An Open-Source Tool for Data Selection in Natural Language Processing", The Prague Bulletin of Mathematical Linguistics, 100, pp. 73-82.

Schwenk, Holger. 2010. "Continuous space language models for statistical machine translation". The Prague Bulletin of Mathematical Linguistics, 93, pp. 137-146.

Schwenk Holger. 2013. "CSLM - a modular open-source continuous space language modeling toolkit", In Interspeech, pp. 1198-1202.

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, pp. 28–35.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010. "Machine Translation Evaluation versus Quality Estimation." Machine Translation, 24(1), pp. 39–50.

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, pp. 710–720.