D7.5: Second report on user group

Philipp Koehn, Germán Sanchis Trilles, Michael Carl

Distribution: Public

CASMACAT
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D7.5

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development.

Project ref no.	ICT-287576
Project acronym	CASMACAT
Project full title	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	October 31, 2014
Actual date of delivery	January 7, 2015
Date of last update	January 7, 2015
Deliverable number	D7.5
Deliverable title	Second report on user group
Type	Report
Status & version	Final
Number of pages	10
Contributing WP(s)	WP7
WP / Task responsible	UEDIN, UPVLC, CBS, CS
Other contributors	
Internal reviewer	
Author(s)	Philipp Koehn, Germán Sanchis Trilles, Michael Carl
EC project officer	Aleksandra Wesolowska
Keywords	

The partners in CASMACAT are:

University of Edinburgh (UEDIN)
Copenhagen Business School (CBS)
Universitat Politècnica de València (UPVLC)
Celer Soluciones (CS)

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

The CASMACAT Project Co-ordinator
Philipp Koehn, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
pkoehn@inf.ed.ac.uk
Phone +44 (131) 650-8287 - Fax +44 (131) 650-6626

Copies of reports and other material can also be accessed via the project's homepage: http://www.casmacat.eu/

Executive Summary

The work plan of the CASMACAT project calls in its Task 7.2 (1 PM) for:

> *Engage early adopters drawn from language service providers and freelance translators. Distribute pre-release versions of the workbench to gather feedback.*

We took advantage of our efforts to gather a user group in previous years as candidates for external field trials of the workbench. We contacted all these users, and we managed to have such collaborations with three of them: DGT (Portuguese Department), Pangeanic, and Autodesk.

Contents

1 Directorate-General for Translation (DGT) — Portuguese Department

The European Commission has a large internal translation department that translates official documents into the 24 official languages of the European Union. Over recent years, statistical machine translation provided by the internal MTEC project, which uses the Moses open source machine translation system, has replaced a legacy translation system. This service is used for glossing foreign language documents and as assistance for the internal translators.

Uptake of machine translation differs by languages, due to machine translation quality and acceptance by translators. The Portuguese department has been for long time enthusiastic about the use of machine translation. The department contacted the CASMACAT project early on, joined the user group, and was interested in providing feedback in an external user study.

1.1 Machine Translation System

The Portuguese Department of the DGT expressed interest in translating documents similar to the publicly available Acquis Communautaire corpus. We used this publicly available corpus as training data and subsampled tune and test sets. The training corpus consists of about 70 million words per language.

The Portuguese Department also provided a preprocessing script especially developed for Portuguese. We used this to train Moses statistical machine translation models using recent settings [1]. The system achieved a BLEU score of 60.3, reflecting the high repetition rate of the corpus, which makes it especially suitable for statistical machine translation. To test out online updating of the translation model, we also built a system that uses the suffix-array based dynamic phrase table [2]. This second system achieved a BLEU score of 60.7, a gain of 0.4. The higher quality of this second system derives from the use of arbitrarily long phrase pairs which come into play in sentence pairs that have strong overlap with the training data.

1.2 CASMACAT Workbench Configuration

In the user study, we were interested in the exploration of interactive machine translation features. After show-casing initial configurations and receiving early feedback, we decided on:

- interactive translation prediction, which uses the *floating prediction* visualization
- display of the translation option array
- use of word alignment information to
  - shade off words in the source sentence that have been already translated
  - highlight the next source word most likely to be translated next
  - orientation of the translation option array to center on the current sentence position
- online updating of the translation model

1.3 Feedback from Users

The workbench was tested informally by two DGT translators, Hilario Fontes and Maria-Cristina De-Preter, who provided us with their personal opinion.

After an initial installation, we received feedback that the workbench is sometimes sluggish, so we made speed improvements, so that they reported that *"now much faster and it is working* and that they are *"liking it a lot"*, especially the *"CasmaCat advanced features"*. They commented that the *"shading of translated source words [...] is really a very interesting feature"*. Unfortunately, the installation only worked on their personal computers, and not at their office, a problem that we could not track down, but may be due to firewall settings.

As feature requests they indicated having count information in the biconcordancer and better handling of tags. They also provided a list of requirements and their current implementation in CASMACAT (the translators also tested the Matecat version of the workbench).

Requirement	**Essential requirement? / Implemented?**
"Traditional" Translation Memories and Concordance	
1. Integration and management (inclusion, deletion and, if possible, ranking) of translation memories (with DGT metadata identifying each segment) for each specific project. 2. Traditional Translation Matches pane. 3. TM pane with the possibility of choosing the number of match suggestions displayed 4. Traditional Concordance pane 5. Concordance with the possibility of choosing the number of matches displayed 6. Possibility of defining a threshold for automatic insertion of MT if no TM match is available.	1. Partially implemented in CasmaCat, and fuller implementation apparently in the pipeline! 2. Yes. Implemented in MateCat 3. This possibility is not essential for this testing? but it would be extremely good to have. It was suggested to MateCat but we don?t know if it is in the pipeline 4. No, but extremely good to have. 5. No, but extremely good to have. 6. It was suggested to MateCat but we don?t know if it is in the pipeline
As you know, in DGT we work with retrieval and reference document tmx files and, for a majority of our projects, we need to know where TM segments come from as we mostly translate legislative documents and therefore we are not "free" to use terminology without taking into consideration what has been used in the reference diplomas.	
Glossaries	
1. Possibility of creating a glossary for the project 2. Possibility of using (downloading to the project and uploading from it) previous glossaries we have	1. Yes. Available in MateCat 2. Yes. Not available yet in MateCat, but apparently in the pipeline.
For some projects, it is very important to have our own glossaries that we can upload and extract from your system	

Requirement	Essential requirement? / Implemented?
Undocking panes	
Possibility of undocking, repositioning, minimizing and maximizing panes (Translation Options and Bi-concordance (CasmaCat), Translation Matches, Concordance and Glossary (MateCat)	Yes. Not available yet in MateCat, but requested and apparently in the pipeline
For Machine Translation, researchers endeavour to make it context sensitive and to have it take into consideration previous and next segment. With the present layout, the screen is occupied with the open segment, almost completely cutting us from the previous and next segments. It is not fair!	
Auto-propagation	
Auto-propagation of segments	No, but very good to have. Available in MateCat
For repetitive documents, it saves a lot of time.	
Document format and creation of translated document(s)	
1. Possibility of using docx (and if possible xlsx documents) 2. Possibility of viewing/previewing translated documents, even if incompletely translated	1. No, but extremely good to have. Available in MateCat 2. Not essential for this testing, but very good to have. Available in MateCat, although incompletely. What MateCat has is, in fact, Pseudo-Translation.
It would not be at all necessary to have the 50 odd fomats MateCat accepts, but it would make it easier for us to be able to import to the project docx and xlsx documents not converted to xliff. It would also allow to preview translated documents without the need for the time-consuming conversion of the xliff file in the native format.	
Sharing projects	
Possibility of having several translators working connected in real-time	Not essential for this kind of testing. Available in MateCat
For testing CasmaCat, this feature is not essential as we can test it with projects that don?t require memory sharing. That is a feature that is being tested in MateCat	

2 Pangeanic

Pangeanic is a translation agency that specializes in the translation of websites, technical documents, and speed multilingual publishing processes of other entities. Pangeanic was one of the pioneers in offering machine translation, both as a service and as a prior step for post-editing in case the translation is required in a short turnaround period. They have been offering both commercially for many years now via their PangeaMT platform.

Having been interested in the new developments in CAT technologies for a long time, they contacted the CASMACAT project during its last year and joined the user group, with the purpose of experimenting with the CASMACAT platform and building their own CASMACAT systems. In particular, they were interested in trying out an English-German (En-De) system and a Japanese-English (Jp-En) system.

2.1 Machine Translation System

Since Pangeanic has already been developing MT systems on their own for some time, they were specially interested in comparing their own systems with the CASMACAT platform with

		English	German	Japanese	English
Training	Sentences	3.1M		2.9M	
	Run. words	40.9M	39.9M	49.4M	36.3M
	Vocabulary	332k	823k	484k	760k
Development	Sentences	1949		1934	
	Run. words	26.1k	25.4k	33.1k	24.4k
	OoV. words	112	274	70	143
Test	Sentences	2000		2000	
	Run. words	28.3k	27.2k	37.3k	27.1k
	OoV. words	154	304	97	184

Table 1: Statistics of the corpora provided by Pangeanic, both En-De and Jp-En. OoV stands for out of vocabulary words with respect to the training data.

the online learning capability enabled. For this purpose, they provided us with the En-De and Jp-En corpora they were currently using for training their own systems, which they had already split into training, development and test. The statistics of such corpora are provided in Table 1.

When evaluated with traditional MT evaluation metrics, the system yielded 45.2 BLEU for the En-De system, and 33.7 BLEU for the Jp-En system.

2.2 CASMACAT Workbench Configuration

When providing Pangeanic with a CASMACAT link for them to test the system, the configuration of the systems provided was as following:

- En-De system with ITP and without online learning capabilities.
- En-De system with ITP and online learning capabilities enabled.
- Jp-En system with ITP and without online learning capabilities.

The purpose of building two En-De systems was to test whether the online learning module was perceived as an advantage by the post-editors. We did not conduct a similar test in the Jp-En case, since the use of Japanese as source language was an experiment per se, since no language with non-Latin alphabet had been tested on the CASMACAT platform until then.

2.3 Feedback from Users

The systems built were tested both by Pangeanic's CEO, Manuel Herranz, who was the one who approached us on the first place, and by one translator for each one of the language pairs. Such feedback is summarized here.

- According to Manuel Herranz, the best part of the CASMACAT platform is provided by the online learning module. This is a great advantage when compared to "static" systems in which the models and weights need to be readjusted in order to include more data, reducing training time and hence optimizing resources. However, he found that, in order to be more useful in the industry, it would be necessary for the CASMACAT platform to include compatibility with commercial formats and XLIFF files. In addition, he thought it would also be beneficial to provide an open API, with the purpose of facilitating its use in commercial situations.

- The En-De translator did not like the fact that the ITP system changed the suffix every time he typed in a word (or letter). He argued that it was uncomfortable for him that, whenever he had decided how to ammend the provided translation, such translation changed as soon as he started typing, and hence he had to think anew how to ammend the new translation provided. He also did not like the fact that "cut and paste" was renedered useless by the ITP interaction scheme. However, in this direction it must be noted that he only used the tool for 15 minutes, which, according to the field trials, is way too few for a translator to get used to the ITP framework and become productive with it. In a more positive note, he also stated that the initial translations provided by the system were good, and even perfect in some cases.
- The Jp-En translator noted that the initial translations were quite good. Interestingly, he did not make any comment abuot the ITP interaction scheme. We feel this is crucial, since we had no experience as of yet with ITP concerning languages without a latin alphabet.

3 Autodesk

Autodesk is a multinational software corporation that develops software for a wide range of industries. Its Geneva offices include a small machine translation research unit, in charge of developing different machine translation engines for the company's use.

Initially, we contacted Mirko Plitt from Autodesk and Ana Guerberof (from Pactera Technology International Ltd.). The first purpose was to build two different CASMACAT systems: one for English-German, and one for English-Chinese. For this purpose, Autodesk would provide us with wordgraphs that would then be fed into the ITP engine for producing the translation proposals. Such systems would then be evaluated at the Pactera offices in Barcelona.

However, the introduction of the wordgraphs from Autodesk into the CASMACAT system was not trivial. There were numerous delays in the communication with Autodesk, and even the head of the MT unit at Autodesk, Mirko Plitt, left the company during the duration of the collaboration, continuing the collaboration with Ventsislav Zhechev. In addition, the wordgraphs could not be integrated directly without any further modification of the CASMACAT platform, since Autodesk had set up different post-processing processes for the translations that such wordgraphs produced. Given that it was not possible to implement such post-process into the CASMACAT workbench, we opted to set up our own post-process for the English-German language pair. However, without having any experience in the processing of Chinese, we opted to leave out the English-Chinese language pair.

3.1 CASMACAT Workbench Configuration

Given that there were numerous difficulties when integrating Autodesk's wordgraphs into the CASMACAT workbench, we opted to provide a single ITP configuration for evaluation, including the following features:

- Confidence measures to highlight potentially incorrect words
- Interactive translation prediction
- Mouse alignments, displaying source (or target) words that are aligned to a certain word when hovering the mouse over it
- Suffix length limitation, shading off suffix words starting from the first potentially incorrect word according to the confidence measures

3.2 Feedback from Users

Initially, it was foreseen that Pactera would evaluate the systems developed. However, after numerous delays introduced by the fact that we had to adapt the CASMACAT pipeline for it to take up the wordgraphs generated by Autodesk, it was not possible within this third year to conduct the study at Pactera's offices. However, we still plan to conduct such study for the sake of research.

4 Univesität Mainz

There was a lively academic exchange between University of Germersheim and CBS, which resulted in a number of short and longer-term visits and scientific collaboration. University of Mainz collected translation process data from more than 30 translation students, annotated and analysed the data in collaboration with CBS, which led to a co-authored book chapter[1]. University of Mainz is also continuing the conference"Translation in Transition: Between Cognition, Computing and Technology"[2] which was first held in Copenhagen (January 30-31, 2014) in connection with CASMACAT , with the plan to turn this into a regular yearly event.

5 Federal University of Minas Gerais

UFMG conducted in collaboration with CBS translation experiments with the CASMACAT workbench using the EMEA corpus. The data is part of the TPR-DB under study name CEMPT13 (see deliverable D1.4). A preliminary evaluation of the datataset is contained in deliverable D1.3, section 6.

6 Members of the User Group

- Autodesk, represented by Mirko Plitt
- Charles University, represented by Ondrej Bojar
- GXP Language Services, represented by Siegfried Armbruster
- European Parliament, represented by Pedro Garcia-Dieguez
- Federal University of Minas Gerais, represented by Fabio Alves
- Pactera Technology Spain, represented by Ana Guerberof
- Pangeanic, represented by Manuel Herranz
- Universitat Autònoma de Barcelona, represented by Anna Aguilar-Amat
- Univesität Mainz, represented by Čulo, Oliver

[1] https://dl.dropboxusercontent.com/u/7757461/07car.pdf
[2] http://bridge.cbs.dk/platform/?q=conference2014

References

[1] Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. Edinburgh's phrase-based machine translation systems for wmt-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 97–104, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[2] Ulrich Germann. Dynamic phrase tables for machine translation in an interactive post-editing scenario. In *Proceedings of the Workshop on Interactive and Adaptive Machine Translation*, pages 20–31, Vancouver, Canada, September 2014. Association for Computational Linguistics.