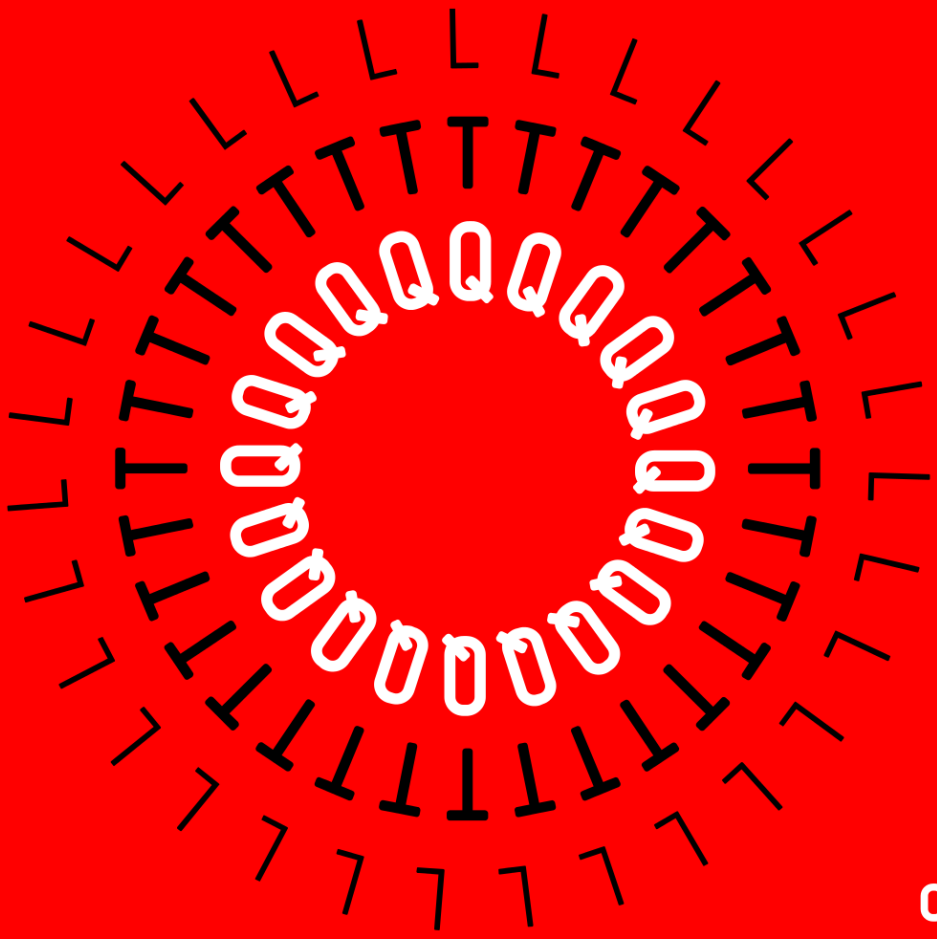


qtleap
quality translation by deep language engineering approaches

Report on the embedding and evaluation of the second MT pilot

DELIVERABLE D3.10

VERSION 1.6 | 2015-11-02

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by
QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by
And supported by the participating institutions:

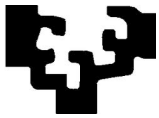

Faculty of Sciences, University of Lisbon


German Research Centre for Artificial Intelligence


Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin


University of Basque Country


University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
1.0	Oct. 20, 2015	Rosa Del Gaudio	HF	First draft
1.1	Oct. 26, 2015	Aljoscha Burchardt	DFKI	Additions and comments
1.2	Oct. 26, 2015	Nora Aranberri	UPV/EHU	Review and minor corrections
1.3	Oct. 27, 2015	Martin Popel	CUNI	Internal review and corrections
1.4	Oct. 27, 2015	António Branco	FCUL	Review and minor corrections
1.5	Oct. 29, 2015	Martin Popel	CUNI	Additions and edits
1.6	Nov. 02, 2015	Aljoscha Burchardt	DFKI	Final edits

Statement of originality
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

REPORT ON THE EMBEDDING AND EVALUATION OF THE SECOND MT PILOT

DOCUMENT QTLEAP-2015-D3.10
EC FP7 PROJECT #610516

DELIVERABLE D3.10

completion
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
ALJOSCHA BURCHARDT (WP3 COORDINATOR)
reviewer
MARTIN POPEL (CUNI)
contributing partners
HF, DFKI, FCUL, UPV/EHU, CUNI

authors
ROSA DEL GAUDIO, ALJOSCHA BURCHARDT, NORA ARANBERRI, ANTÓNIO BRANCO, MARTIN POPEL

Contents

1 Introduction

In this document, we present the result of extrinsic evaluation of the QTLeap Machine Translation (MT) Pilot 2, compared with Pilot 0 and Pilot 1. In QTLeap, the MT development is structured by pilot engines that serve as a kind of milestone in the creation of MT engines for the seven project languages. While Pilot 0 served an SMT baseline (trained on domain corpora if available), Pilot 1 was a first "deeper" MT system including more linguistic and knowledge-driven components. Pilot2 is the result of experimenting with and inclusion of lexical knowledge (such as WSD) into the MT pipelines. More details on the Pilots can be found in project deliverables D2.4 and D2.8.

As MT evaluation is notoriously difficult and can be resource intensive, QTLeap makes use of a mix of several intrinsic and extrinsic evaluation procedures to track improvements and at the same time get feedback and inspiration for further improvements of the MT engines. While the intrinsic evaluation of Pilot2 (against reference translations with automatic measures and using manual error annotation) is documented in D2.8, the present deliverable is concerned with the extrinsic (or user) evaluation.

This extrinsic evaluation is based on the integration of MT services in PcMedic Wizard, an online helpdesk application developed by the industrial project partner HF as part of its business. This evaluation is a follow up of the evaluation carried out for Pilot 0 and Pilot 1, reported respectively in D3.6 and D3.8. The focus of this evaluation is to assess the added value of the translations in terms of their impact on the performance of the QA system of the helpdesk.

In the present evaluation, the focus is to compare the impact of the translation delivered by Pilot2 with the impact of the translations delivered by Pilot 0 and Pilot 1. As for previous evaluation, the present evaluation is composed of two distinctive parts. The first part focuses on evaluating how the translation affects the answer retrieval component of the question and answer (QA) algorithm. The second part focuses on outbound translation, aiming to evaluate to what extent it delivers a clear and understandable answer to final customers without the intervention of a human operator.

The evaluations have been carried out using an online platform designed by HF for this purpose. The testing subjects have been volunteers recruited by project partners that match the profile of the typical HF users as closely as possible (non-experts, mixed in age, etc.). As it would have gone far beyond the limits of the project to build a full simulation of a repair situation, e.g., in a laboratory with modified, malfunctioning equipment, this user evaluation measures perceived usefulness of MT when integrated into the HF business scenario.

One note on the authors of this Deliverable. After Martin Popel served as an internal reviewer to the draft, he added some new ideas and elaborated on existing content to this Deliverable so that we decided to add him as an author to acknowledge this.

2 Embedding of Pilot 2

The embedding of MT Pilot 2 was performed along the same lines that were followed for the embedding of MT Pilot 1, described in deliverable D3.8, and are briefly indicated again in the subsections below.

2.1 The PcMedic Wizard Application

The PcMedic Wizard application developed by HF offers technical support service by chat. Technical support can usually be divided into three levels: first-level (front line), second-level, and third-level. Most of the users' requests for help are straightforward and simple, and can be easily handled by the first-level operator. Literature has shown that the majority of user requests can be answered at this level, as they are "simple and routine" and do not require specialized knowledge Leung and Lau [2007]. At the same time, these kinds of requests represent the majority of the total requests and are responsible for long wait times, leading to user dissatisfaction.

The PcMedic Wizard application attempts to address this specific context, trying to automate the process of answering first-level user requests. The area of specialization of this service is basic computer and IT troubleshooting for both hardware and software. The process of providing support to end-users involves remote written interaction via chat channels through a call center. This process of problem solving can be made efficient by a Question Answering (QA) application that helps call center operators prepare replies for clients.

Using techniques based on natural language processing, each query for help is matched against a memory of previous questions and answers and a list of possible replies from the repository is displayed, ranked by relevance according to the internal heuristics of the support system. If the top reply scores above a certain threshold, it is automatically returned to the client. If no reply score overs the threshold, the operator is presented with the list of possible answers delivered by the system and he can (a) pick the most appropriate reply, (b) modify one of the replies, or (c) write a completely new reply. In the last two cases, the new reply is used to further improve the QA memory.

As there are currently no multilingual operators present at HF, QTLeap evaluates the usefulness of returned answers relative to a known reference from the database, thus limiting itself to the fully automatic case. This is enforced as the questions that are used in the evaluation described below are provided by the system.

2.2 The embedding the MT services

The PcMedic Wizard application is implemented in a proprietary xRM platform (Extended Relationship Management) developed by HF that is called WhiteBox. This platform incorporates a set of different features supporting different types of activities and data, from managing remote and on-site technical supports, laboratory activities, to integrating complex information such as human resource, different departments, suppliers, etc.

This platform is based on the integration of different technologies such as VoIP systems, Flex, SignalR, interconnecting social networks and several web services variants. Its modular design makes the integration of new services relatively easy.

Regarding the integration of the MT Pilot 2 systems (described in deliverable D2.8), some adjustments were carried out mostly at the level of database design supporting the multilingual information introduced by the translations.

The embedding of the actual translation services was obtained by using the web-services (see Task 3.2) based on the documentation provided in full detail in deliverable D3.4.

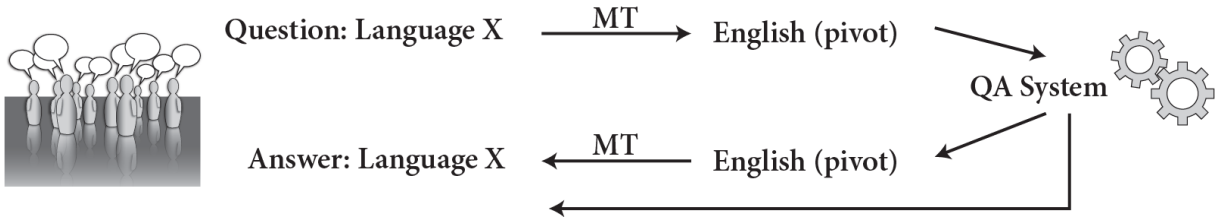


Figure 1: The workflow with the MT services

3 The Evaluation's main objective

The main objective of this evaluation is to assess the performance of Pilot2 MT systems in the PcMedic Wizard system. The main challenge was to separate the evaluation of the system as a whole from the evaluation of the MT component. The intention was not to assess the quality of the PcMedic Wizard application, but rather to assess the impact of the MT services on the application. The evaluation design was conceived keeping in mind this issue. Figure 1 shows PcMedic Wizard's workflow with the embedded MT services. There are two distinct places where MT services are used in the application. The first time occurs when the incoming user request is translated from the original language to English, the pivot language in which the data is stored in the repository. This translation is used by the QA search algorithm to retrieve a possible answer.

After an answer is found in English, MT services are called to translate the answer back to the user's original language.

This means that the MT services interact with the system in two different stages and for two very different purposes. In the first, inbound retrieval step, the translation is not presented to a human, but it is only used by an algorithm. By contrast, in the second, outbound presentation step the translation is presented to the final user. This evaluation consists of two different parts, focusing on each of these steps separately.

4 Evaluating the retrieval step

The main idea of this step is to compare the result obtained when the original English question is used by the QA algorithm with the result obtained when the question is translated from a different language to English by the MT services.[1] This evaluation is possible because all of the language versions of questions and the answers share a common ID in the database and the algorithm works with English as pivot language.

As the QA system currently works only with English/Portuguese, there is no baseline for the other languages. Accordingly, the performance of cross-lingual answer retrieval has been evaluated against the English reference answer(s). If the translation is appropriate for this kind of task, the QA heuristic should retrieve the same answer for each question as it does in the English experiment. We have also compared the list of answers delivered for each question along with their confidence scores. In this way, it was possible to measure how far the results returned for a translated question are from the results of using the original English question.

[1]As German is not part of QTLeap WP5 where lexical-semantic resources have been developed, developments of Pilot2 for German have been concentrated on the more important and challenging outbound translation direction using an existing WSD system for English. Therefore, German is not part of the evaluation of the retrieval step for Pilot2.

As reported in D3.1 and D3.3, the QTLeap corpus composed of 4000 question&answer pairs was collected and human translated by HF in order to support the development and the evaluation of the MT services. This corpus also represents a gold standard for the translations. In order to be used for developing and testing the MT systems, it was divided into four batches. In the current evaluation, we used the third batch (called Batch3). This subcorpus was never used, neither for developing any of the pilots or for evaluation. All the 1000 questions from this third batch were used to automatically assess the performance of the retrieval step. The database used for testing this step included all 4000 interactions from the entire corpus in English.

Each question for each language was translated to English using the MT service and then given to the QA algorithm. The list of results obtained has been compared with the results obtained when the original English question is used.

The Pilot 0 results for Spanish reported in this deliverable are Pilot 0-comparable, that is Pilot 0 trained on Europarl only, so it can be fairly compared with Pilot 1 and Pilot 2, which are also trained on Europarl only.

4.1 Recall

The QA system produces a list of candidate answers with a confidence score ranging between 1 and 100, where 100 means that the QA search module is quite sure that the answer is correct for the given question. The score represents the confidence of the algorithm based on several factors, such as lexical similarity, how many times a given answer was used, and when it was used the last time. Even if a new question is very similar or equal to a question in the database, if the associated answer was used just once several months ago, the final score will be lower.

Following the PcMedic Wizard workflow, the answer is displayed to the client without human intervention only if there is an answer with a confidence score above 95 points. This is the preferred situation and the goal of the QA system, as it saves time and money in the long term. If no answers are retrieved with a confidence score above 95, the top five results with a confidence score above 75 are shown to an operator, who can choose to adopt one of the answers (with our without changes), or accept none of them. For HF, this option is less preferred as it saves only some time/money. If no answer scores above 75 points, the question will be answered by an operator with no help from the system. (Because of the penalization of old and infrequent answers, it sometimes will happen that a correct answer may receive a score below the thresholds of 95 or 75.)

As the language in the QA system is English and the heuristic is tuned to work with this language, the percentage of answers obtained in this way to our test set represents the upper bound of the actual system.

The next three tables present the percentage of how many questions the QA algorithm is able to find a candidate answer for within a certain confidence score interval using the Pilot 0 (Table 1), Pilot 1 (Table 2) and Pilot 2 (Table 3).

English, representing the upper bound of the system, is highlighted in gray. The average for all the languages, except English, is also presented in the last column. This value represents the overall performance of the system when MT services are used. Note that it is not the intention to compare the performance of pilots across languages as that would require creating “equal conditions” – whatever this may mean in practice given the diversity of resources available for the different languages. Within this project and evaluation, the goal is to compare the performance of pilots within a given language pair.

Score	EN	EU	BG	CS	NL	PT	ES	avg.
>=95	70.30%	11.66%	16.32%	21.02%	19.96%	15.23%	17.82%	14.57%
75-94	26.70%	16.28%	25.83%	25.53%	28.69%	26.15%	28.03%	21.50%
50-74	2.90%	46.13%	44.64%	41.74%	38.72%	44.19%	41.54%	36.71%
25-49	0.10%	23.12%	12.51%	11.21%	12.04%	13.83%	12.11%	12.12%
1-24	0.00%	2.81%	0.70%	0.50%	0.60%	0.60%	0.50%	0.82%

Table 1: Percentage of the answers delivered by QA System and their scores for Pilot 0

Score	EN	EU	BG	CS	NL	PT	ES	avg.
>=95	70.30%	9.14%	19.72%	21.74%	18.57%	11.72%	11.04%	13.13%
75-94	26.70%	10.74%	27.03%	30.66%	27.31%	24.15%	23.39%	20.47%
50-74	2.90%	47.79%	39.74%	38.58%	40.06%	46.69%	45.48%	36.91%
25-49	0.10%	30.52%	12.61%	8.12%	13.35%	16.73%	19.18%	14.36%
1-24	0.00%	1.81%	0.90%	0.90%	0.70%	0.70%	0.90%	0.85%

Table 2: Percentage of the answers delivered by QA System and their scores for Pilot 1

In general, when no translation is used, the English QA system can automatically answer a question without human intervention in 70% of the cases. In 27% it provides help for operators supporting them in finding the right answer. Only in about 3% of the cases the operator is left without any help.

When the translation services are used, the number of answers scoring 95 or more drop considerably in all the three pilots. However, there is an improvement in the performance from Pilot 0 to Pilot 2 for almost all the languages.

For Basque, the percentage of answers with a score of 95 or more decreases slightly From Pilot 0 to Pilot 2. But if we consider the scores above 75%, that is, segments that are used by the PcMedic Wizard, we see that Pilot 2 has outperformed Pilot 0 (Pilot 0 27.94% vs. Pilot 2 28.73).

For Bulgarian, we can observe an improvement for both scores when comparing Pilot 0 with Pilot 1. However, when looking at Pilot 2 results, there is a small decrease in performance for the answers with a score above 95, and a small increase for the answers with a score between 75 and 94. Two languages, Czech and Spanish present improvements from Pilot 0 to Pilot 1 and from Pilot 1 and Pilot 2, while for Dutch and Portuguese, performance decreases from Pilot 0 to Pilot 1. Overall, the results of Pilot 2 outperform results obtained with Pilot 0.

4.2 Precision

While the recall indicates where the QA system can deliver a suggestion for a given question, it does not indicate the quality of this suggestion in relation to the gold standard, i.e. whether the high-confidence suggestion is the correct suggestion.

The following three Tables show the percentage of cases in which the first answer of the gold standard appears in the list of the answers obtained using the translated question, particularly if it appears in the first place, in the first two or in the first three places.

It also shows the average score of the answers for each position. In the first case, for example, the score shows the average score for all answers corresponding to the first English answer and appearing in the first place. In the second setting, the score shows the average score for all answers corresponding to the first English answer and appearing

Score	EN	EU	BG	CS	NL	PT	ES	avg.
>=95	70.30%	11.51%	19.32%	24.40%	20.42%	17.80%	20.66%	16.30%
75-94	26.70%	17.22%	27.33%	33.70%	30.03%	27.20%	28.08%	23.37%
50-74	2.90%	52.45%	40.54%	35.10%	39.64%	43.70%	39.92%	35.91%
25-49	0.10%	18.32%	12.11%	6.40%	9.41%	10.90%	11.13%	9.75%
1-24	0.00%	0.50%	0.70%	0.40%	0.50%	0.40%	0.20%	0.39%

Table 3: Percentage of the answers delivered by QA System and their scores for Pilot 2

in first or second place.

English scores are constant, because English represents the gold standard and thus, it always represents the first best answer and its score.

Score	EN	EU	BG	CS	NL	PT	ES	avg.
First (%)	100.00%	44.70%	61.90%	65.90%	60.60%	55.50%	59.80%	58.00%
Score	95.46	69.27	74.64	76.54	75.58	73.75	74.68	74.08
First 2 (%)	100.00%	57.30%	75.30%	77.70%	72.80%	67.60%	71.70%	70.00%
Score	95.46	65.9	71.49	74.01	73.12	70.91	72.09	71.25
First 3(%)	100.00%	62.80%	79.20%	82.60%	77.70%	72.80%	77.10%	75.00%
Score	95.46	64.26	70.37	72.8	72.05	69.55	70.59	69.94

Table 4: Percentage of answers delivered as first candidates for both English and target language with Pilot 0

Score	EN	EU	BG	CS	NL	PT	ES	avg.
First (%)	100.00%	36.00%	61.10%	66.90%	57.60%	52.40%	51.90%	54.00%
Score	95.46	64.13	76.12	78.25	74.82	70.53	69.26	72.18
First 2 (%)	100.00%	48.50%	73.30%	80.70%	71.10%	65.20%	64.50%	67.00%
Score	95.46	60.75	73.04	74.85	71.56	67.87	66.5	69.09
First 3 (%)	100.00%	55.30%	77.50%	85.50%	76.20%	71.30%	69.70%	73.00%
Score	95.46	58.56	72.01	73.57	70.38	66.42	65.15	67.68

Table 5: Percentage of answers delivered as first candidates for both English and target language with Pilot 1

Overall, these data indicate that in about 60% of the cases the first answer suggested by the QA system corresponds to the first suggestion of the gold standard when Pilot 2 MT is used. If the comparison is broadened to check if the first suggestion of the gold standard appears in the first two or three places of the MT answers, this percentage increases to 73% and to 78% when considering the average of Pilot 2.

In this case, only Bulgarian does not present any improvement along the three pilots, obtaining the best performance with Pilot 0. For four languages, namely Basque, Dutch, Portuguese and Spanish, Pilot 0 outperforms Pilot 1, but then Pilot 2 gets better results then the previous two pilots. Only Czech presents a constant improvement along the three pilots. This means that for five languages out of six, Pilot 2 performs better than the previous pilots.

Score	EN	EU	BG	CS	NL	PT	ES	avg.
First (%)	100.00%	48.20%	61.50%	70.30%	63.90%	58.80%	60.10%	60.00%
Score	95.46	69.12	76.27	80.45	76.51	74.82	76.95	75.69
First 2(%)	100.00%	59.50%	73.00%	82.30%	76.40%	70.90%	73.80%	73.00%
Score	95.46	65.92	73.25	77.66	74.19	71.87	73.69	72.76
First 3(%)	100.00%	67.00%	77.30%	86.60%	80.00%	76.80%	79.10%	78.00%
Score	95.46	63.59	72.13	76.6	73.24	70.22	72.38	71.36

Table 6: Percentage of answers delivered as first candidates for both English and target language with Pilot 1

5 Evaluating the publication step

In the design of this evaluation multiple issues had to be taken in consideration. First, as the PcMedic Wizard application is currently used only in Portugal for the Portuguese language, no real baseline to evaluate the business case in a multilingual scenario was available.

Second, for the same reason, there was no access to real (non-Portuguese) users with real questions about the software or hardware they are using. As a result, the evaluation had to approximate the real context.

Third, the project had to rely on volunteer evaluators found by the partners, which made it difficult to obtain evaluators that match the typical HF customer (low computer proficiency, all ages, all types of educational background, etc.).

The evaluation set-up of Pilot 2 follows closely the evaluation of Pilot 1, where evaluators were asked to compare the translations of Pilot 0 and Pilot 1. For evaluating Pilot 2, we asked evaluators to rank the three different translations delivered by Pilot 0, Pilot 1 and Pilot 2.

In the main evaluation page, the evaluators were presented with the reference answer and the translations of the three Pilots (anonymized as “A”, “B” and “C”, in random order, so the evaluation is blind). They were asked to rank these three alternative answers against the reference answer. The precise instructions to the annotators were (presented in their mother tongue):

> Read the following three alternative answers and rate them from best to worst.
>
> If you think two answers have the same quality, you can assign the same number twice or more.
>
> For example, you can rate answers A-B-C as 1-2-3 or 2-1-3 or 2-2-1 or 1-1-1 or any other combination of these numbers that you find appropriate.

As will become clear below, this boils down to a ranking task between the three pilots, which is common practice in human MT evaluation, e.g., as performed in the WMT Shared Tasks.

All partners helped in this evaluation by translating the evaluation interface and recruiting volunteers for the evaluation. The selection of these volunteers tried to take into account the real user profile of people who use computers in their everyday lives, but who are not experts.

Figure 2: Evaluation Interface: Selecting a Language

5.1 Experimental setup

A new web-based interface was set up and translated for all the languages. This interface for the extrinsic evaluation of Pilot 2 is available at `http://83.240.145.199/questionnaire/pilot2/`.

On the first page (Figure 2), evaluators select their language.

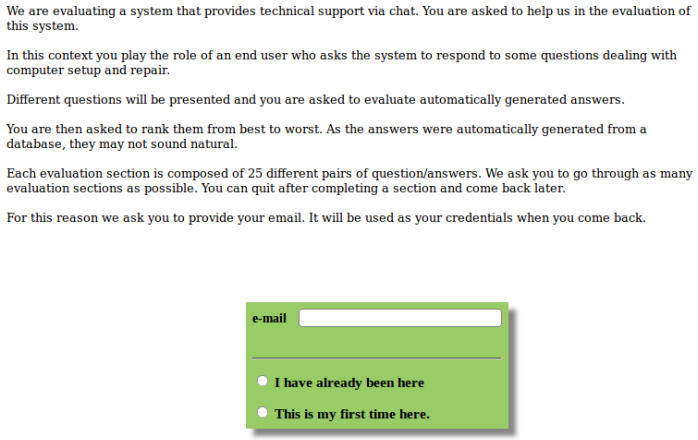


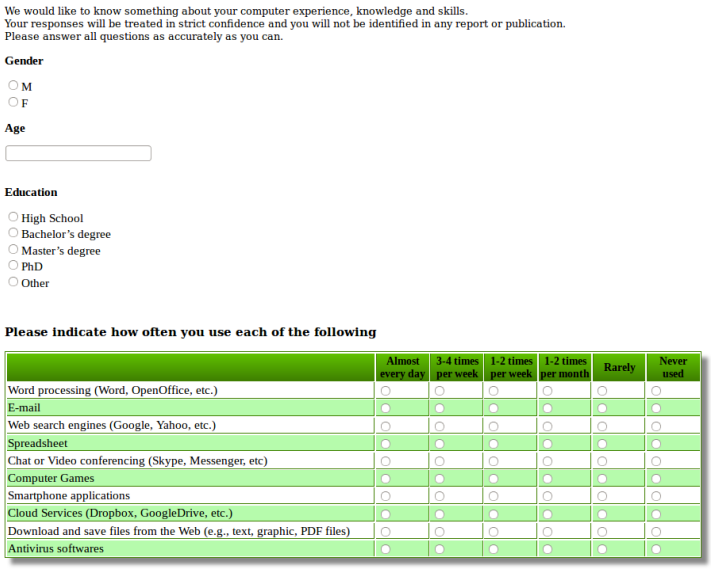
Figure 3: Evaluation Interface: Introduction and log in

The second page (Figure 3) provides a brief explanation of the evaluation task. The evaluators provide their e-mail to register in the system. Registration allows them to quit the evaluation at any time and come back later without losing the evaluation work they have already done and also helps ensure that, in the next turn, they will be presented only with interactions that still need to be evaluated.

When an evaluator registers with the system, he or she is asked for basic information about age, sex, education, and familiarity with information technology (Figure 4).

The next page presents Form 1 (Figure 5), and represents the start of the evaluation. Here a question in the selected language is presented and the evaluator is asked to provide a self-estimation of his knowledge level (high, medium, low) about the subject involved in the question.

In Form 2 (Figure 6), the same question is presented, followed by the manually translated answer and by the three answers generated by Pilot 0, Pilot 1 and Pilot 2 (in randomized order). The subject is asked to read the reference answer and the three fol-



We would like to know something about your computer experience, knowledge and skills.
Your responses will be treated in strict confidence and you will not be identified in any report or publication.
Please answer all questions as accurately as you can.

Gender

M
F

Age

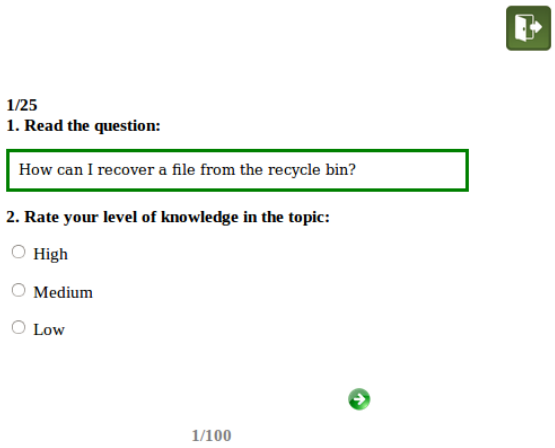
Education

High School
Bachelor's degree
Master's degree
PhD
Other

Please indicate how often you use each of the following

	Almost every day	3-4 times per week	1-2 times per week	1-2 times per month	Rarely	Never used
Word processing (Word, OpenOffice, etc.)						
E-mail						
Web search engines (Google, Yahoo, etc.)						
Spreadsheet						
Chat or Video conferencing (Skype, Messenger, etc)						
Computer Games						
Smartphone applications						
Cloud Services (Dropbox, GoogleDrive, etc.)						
Download and save files from the Web (e.g., text, graphic, PDF files)						
Antivirus softwares						

Figure 4: Evaluation Interface: Basic information



1/25
1. Read the question:

How can I recover a file from the recycle bin?

2. Rate your level of knowledge in the topic:

High
Medium
Low

1/100

Figure 5: Evaluation Interface: Form 1

lowing answers. Then, being told to suppose that the reference answer is correct, the evaluator is asked to rank the three answers from best to worst. It is possible to assign the same rank to more than one answer.

This evaluation was carried out in a controlled setting in order to avoid having to deal with different variables that interfere with the real objective of this evaluation, such as having a relatively small multilingual database and no previous data on a multilingual scenario. Furthermore, a direct field test would lead to the problem that the questions would differ between evaluations and complicate comparison of the results.

For these reasons 100 question/answer pairs from QTLeap Corpus batch 3 were used. Each project partner recruited volunteers that were not IT experts in order to simulate the typical user of the PcMedic Wizard application. The same 100 interactions have been evaluated for all language pairs. All the question/answer pairs were evaluated at least by 2 volunteers for each language, with a global average of 2.8 evaluations per interaction.

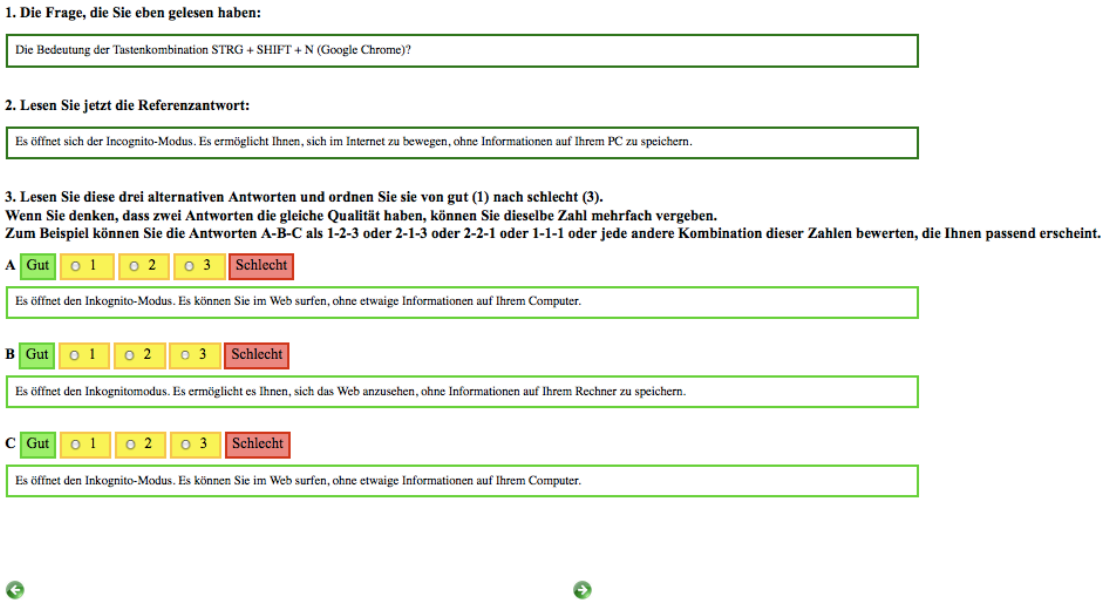


Figure 6: Evaluation Interface: Form 2

5.2 Result and Discussion

This section presents the results of the evaluation and the inter-annotator agreement.

5.2.1 Manual evaluation of translations from English

Table 7 shows the average score obtained with the ranking, where 1 means best and 3 worst. This table offers a first insight in the performance of the different pilots. For all the languages, Pilot 1 answers obtain a worse score than Pilot 0. For Basque, Pilot 2 shows an improvement over Pilot 1, but not over Pilot 0. For Bulgarian, the best results are obtained by Pilot 0, followed by Pilot 1. For the remaining five languages, Pilot 2 outperforms both Pilot 1 and Pilot 0.

	EU	BG	CS	NL	DE	PT	ES
Pilot 0	1.31	1.94	2.17	1.87	2.35	2.33	2.15
Pilot 1	2.27	2.09	2.22	2.03	2.39	2.42	2.36
Pilot 2	1.93	2.12	2.07	1.81	2.33	2.18	1.78

Table 7: Average score for the three pilots (1 best, 3 worst)

Table 8 show the same data presented in Table7 after their normalization. It means that, for example, if an interaction was evaluated as 1-3-3, it was normalized to 1-2-2, or if the evaluation was 3-3-3, it was normalized to 1-1-1.

	EU	BG	CS	NL	DE	PT	ES
Pilot 0	1.23	1.33	1.53	1.82	1.40	1.68	1.84
Pilot 1	2.16	1.48	1.55	1.97	1.41	1.75	2.03
Pilot 2	1.83	1.49	1.40	1.77	1.35	1.52	1.48

Table 8: Average score for the three pilots (1 best, 3 worst) using normalized data

Table 9 shows more detailed information on the performance of Pilot 2 in comparison with the other two pilots. The first row presents the percentage of how many times Pilot 2 translations were ranked above both Pilot 1 and Pilot 0. The second and third row show the percentage of cases where Pilot 2 obtained the same rank as one of the two pilots and better than the one. The next row accounts for the cases where the three pilots got the same rank. Finally, the last row sums up the results of previous rows and reports on how often Pilot 2 translations were ranked equal or above the other two pilots. For 5 languages, namely Bulgarian, Czech, German, Portuguese and Spanish, Pilot 2 performs better or has the same performance than the other two pilots. The language that present the best results is Spanish with 71.97%, with 53.98% of translations ranked above both Pilot 1 and Pilot 0.

	EU	**BG**	**CS**	**NL**	**DE**	**PT**	**ES**
P2 better than P1 and P0 (%)	12.68	7.18	12.00	26.83	26.83	31.93	53.98
P2 equal to P1 and better than P0 (%)	1.09	5.09	23.00	13.17	2.44	7.83	5.54
P2 equal to P0 and better than P1 (%)	6.88	9.72	5.00	1.46	4.88	5.72	5.19
P2 equal to P1 and P0 (%)	11.59	36.34	23.50	3.41	37.56	15.96	7.27
Total	32.24	58.33	63.50	44.87	71.71	61.44	71.98

Table 9: Comparison between pilots: when Pilot 2 performs better

Table 10 and Table 11 show the comparison between Pilot 2 and Pilot 0, and between Pilot 2 and Pilot 1, respectively. Let's focus now on Table 10. As we can see in row *c*, the percentage of ties differs across languages: the Dutch Pilot 2 was evaluated as equal to Pilot 0 only in 5.65%, while for Bulgarian it was in 58.10% of the evaluations. Therefore, we cannot compare the relative quality of Pilot 2 and Pilot 0 only based on the number of cases when Pilot 2 was judged strictly better than Pilot 0 (row *a*), or only based on the number of cases when Pilot 2 was judged better or same as Pilot 0 (row *c*). Thus, in the last row *e*, we report the percentage of non-tying comparisons where Pilot 2 was judged better than Pilot 0, that is, *P2 better ignoring ties than P0 (%)* equals

$$\frac{\#(P2 \text{ better than } P0)}{\#(P2 \text{ better than } P0) + \#(P2 \text{ worse than } P0)} \times 100\%$$

	EU	**BG**	**CS**	**NL**	**DE**	**PT**	**ES**
a) P2 better than P0 (%)	15.22	13.89	37.00	46.09	29.27	46.08	62.29
b) P2 worse than P0 (%)	63.04	28.01	31.50	48.26	27.76	27.11	22.49
c) P2 equal to P0 (%)	21.74	58.10	31.50	5.65	42.93	26.81	15.22
d) P2 better or same as P0 (%)	36.96	71.99	68.50	51.74	72.20	72.89	77.51
e) P2 better ignoring ties (%)	26.09	42.21	52.75	48.92	51.47	59.49	69.90

Table 10: Comparison between Pilot 2 and Pilot 0

Figures 7 and 8 provide a graphical representation of Tables 10 and 11, respectively. The languages (vertical bars) in the figures are sorted by the *better ignoring ties* scores, which are plotted as a dark blue line. We can see that for four languages, Pilot 2 is better than the respective Pilot 0 (the *better ignoring ties* score is higher than 50%). Also, all languages except for Basque have Pilot 2 at least as good as the respective Pilot 0 (the

	EU	BG	CS	NL	DE	PT	ES
a) P2 better than P1 (%)	52.54	22.69	21.50	45.22	32.20	45.48	71.98
b) P2 worse than P1 (%)	11.59	25.00	10.50	28.70	23.85	15.36	8.30
c) P2 equal to P1 (%)	35.87	52.31	68.00	26.09	43.09	39.16	19.72
d) P2 better or same as P1 (%)	88.41	75.00	89.50	71.30	76.10	84.64	91.70
e) P2 better ignoring ties (%)	70.48	47.97	55.50	58.26	54.89	65.06	81.83

Table 11: Comparison between Pilot 2 and Pilot 1

yellow bar reaches over 50%). From Figure 8 we can see that for all languages except for Bulgarian, Pilot 2 is better than Pilot 1.

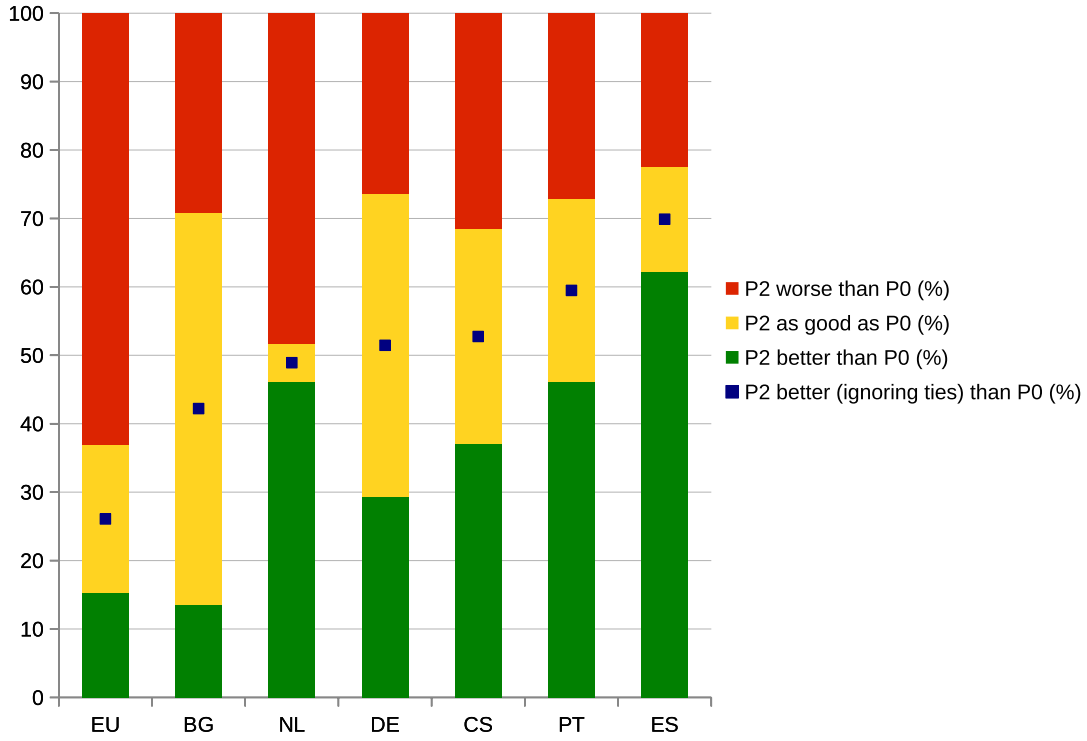


Figure 7: Comparison of Pilot 2 and Pilot 0, breakdown of the human evaluation

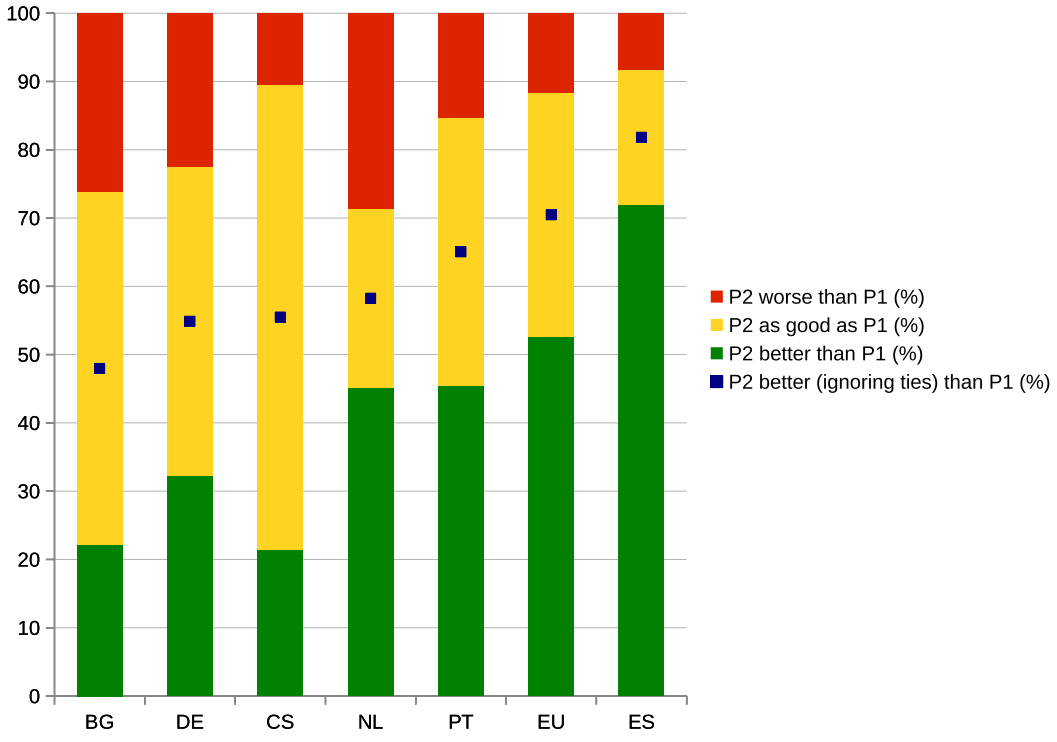


Figure 8: Comparison of Pilot 2 and Pilot 1, breakdown of the human evaluation

5.2.2 Inter-annotator agreement

Table 12 show the results for inter-annotator in terms of accuracy for each language.

Language	Accuracy
Basque	0.47
Bulgarian	0.56
Czech	0.64
Dutch	0.58
German	0.61
Portuguese	0.46
Spanish	0.61

Table 12: Inter-annotator agreement

The table presents accuracy in $[0, 1]$, that is the portion of cases where two evaluators ranked two systems in the same way (*better*, *worse*, or *tie*), computed as $C/(C + D)$ where: C is the number of concordant pairs and D is the number of discordant pairs. A ranking pair is here considered concordant, only if both evaluators agree; all other pairs (e.g. *tie–better*) are considered discordant.

As expected from experience in many tasks related to human judgments on translation quality, inter-annotator agreement is not particularly high. It seems that a lot of subjectivity and personal taste is involved even in such a comparably straightforward ranking task. Yet, from the perspective of the HF company, this variety in "user satisfaction" about a particular advice given is part of the business that cannot be controlled.

5.2.3 Correlation of manual and automatic evaluation

We finally compared this human extrinsic evaluation data with the automatic intrinsic performance measure BLEU. Table 13 shows the BLEU scores of the three pilots as described in D2.8 and the difference between Pilot 2 and the other two Pilots.

	EU	BG	CS	NL	DE	PT	ES
P0 BLEU	18.59	17.72	21.34	25.98	34.82	13.75	16.23
P1 BLEU	9.62	16.36	20.44	18.15	31.56	12.86	10.73
P2 BLEU	11.27	16.91	21.89	19.66	29.57	15.51	24.32
BLEU(P2)−BLEU(P0)	−7.32	−0.81	0.55	−6.32	−5.25	1.76	8.09
BLEU(P2)−BLEU(P1)	1.65	0.55	1.45	1.51	−1.99	2.65	13.59

Table 13: Comparison between all Pilots in terms of BLEU on QTLeap Batch3a (see D2.8).

Figures 9 and 10 present the BLEU difference (dark red bars) in relation to the difference between Pilots according to the human extrinsic evaluation (violet bars).

For this purpose, we scaled the *better ignoring ties* score (defined in Section 5.2.1) to the same range as BLEU difference, that is $\langle -100; +100 \rangle$, which boils down to %(P2 better than PX) - %(P2 worse than PX), where PX means Pilot 0 (in Figure 9) or Pilot 1 (in Figure 10). The languages (bars) in Figures 9 and 10 are sorted according to this human evaluation (that is, in the same order as in Figures 7 and 8, respectively.)

It is interesting that BLEU differences almost always agree with the user ratings on the comparison of two systems. There are just three exception: German Pilot 2 vs.

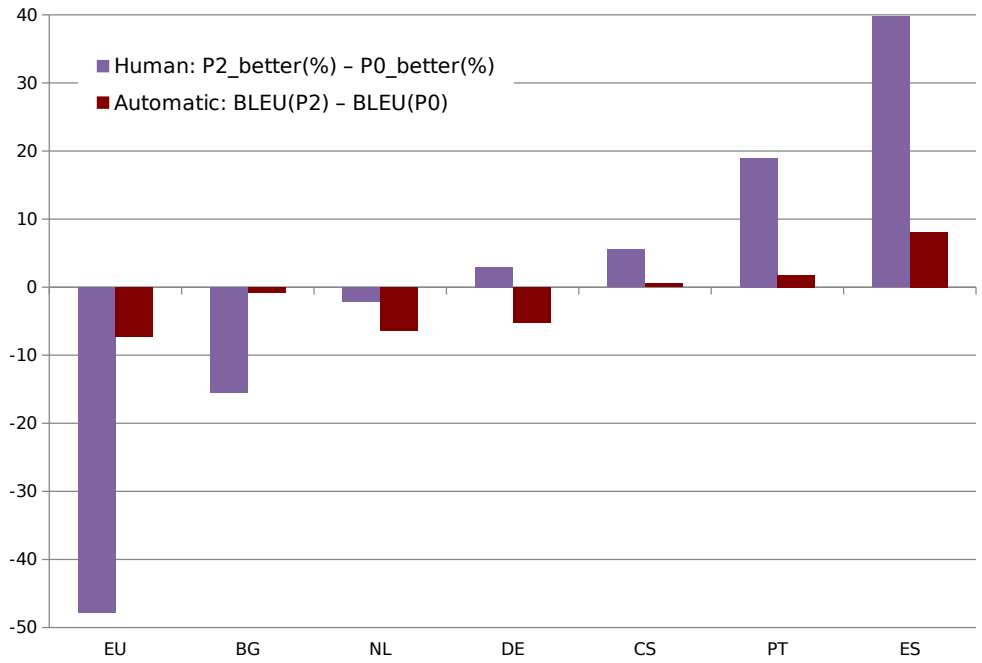


Figure 9: Comparison of user evaluation results and BLEU scores for Pilot 2 and Pilot 0

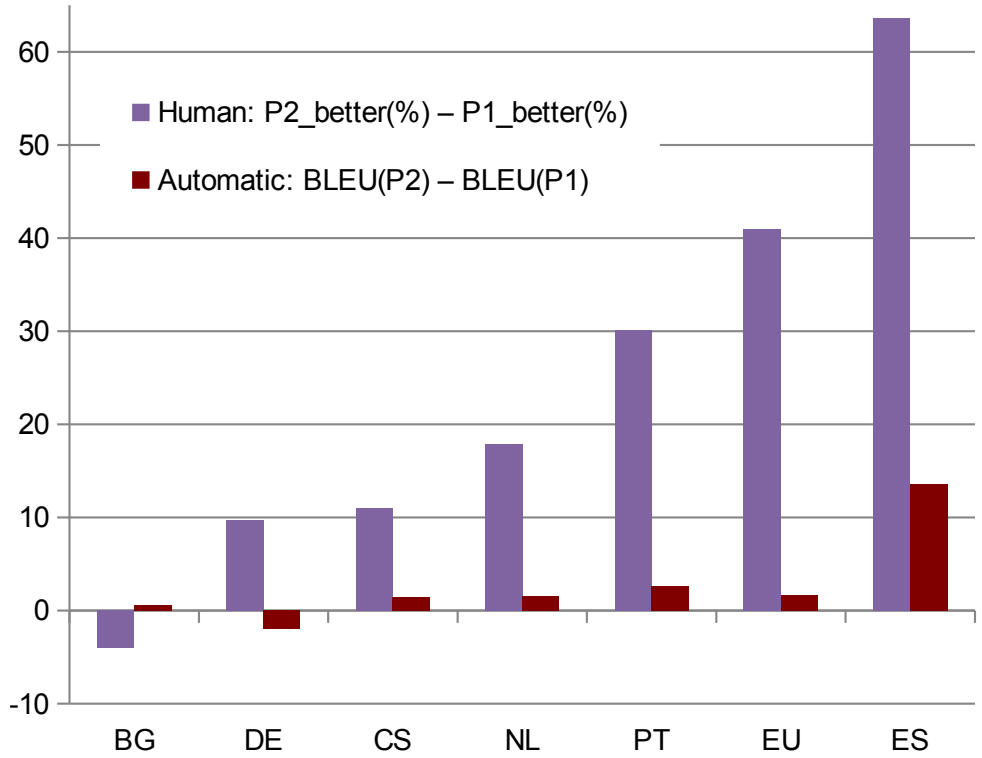


Figure 10: Comparison of user evaluation results and BLEU scores for Pilot 2 and Pilot 1

Pilot 0 (see Figure 9), and German and Bulgarian Pilot 2 vs. Pilot 1 (see Figure 10). For German, the user ratings are generally more in favor of Pilot 2 (unlike BLEU).

6 Conclusion

This extrinsic evaluation of Machine Translation systems created in QTLeap has compared the performance of three different systems for each one of the seven languages. While the Pilot 0 systems are (Moses) baselines, the Pilot 1 systems make use of "deeper" linguistic information, albeit still in an entry-level state. Pilot 2 represents an evolution of Pilot 1, enhanced with lexical semantics. As machine translation evaluation is notoriously difficult, the project has performed both an intrinsic evaluation (reported in deliverable D2.8) and the extrinsic evaluation reported here.

In the extrinsic evaluation at hand, the MT systems have been tested by volunteer subjects in a usage scenario of project partner HF, namely a chat-based PC helpdesk scenario (PcMedic Wizard). The evaluation has shown that the Pilot 2 systems are already performing better than or similar to the first two pilots for almost all the project languages.

This performance is an achievement and is encouraging given the complexity and known issues when integrating semantic information.

References

Nelson K. Y. Leung and Sim Kim Lau. Information technology help desk survey: To identify the classification of simple and routine enquiries. *Journal of Computer Information Systems*, 47(4):70–81, 2007.