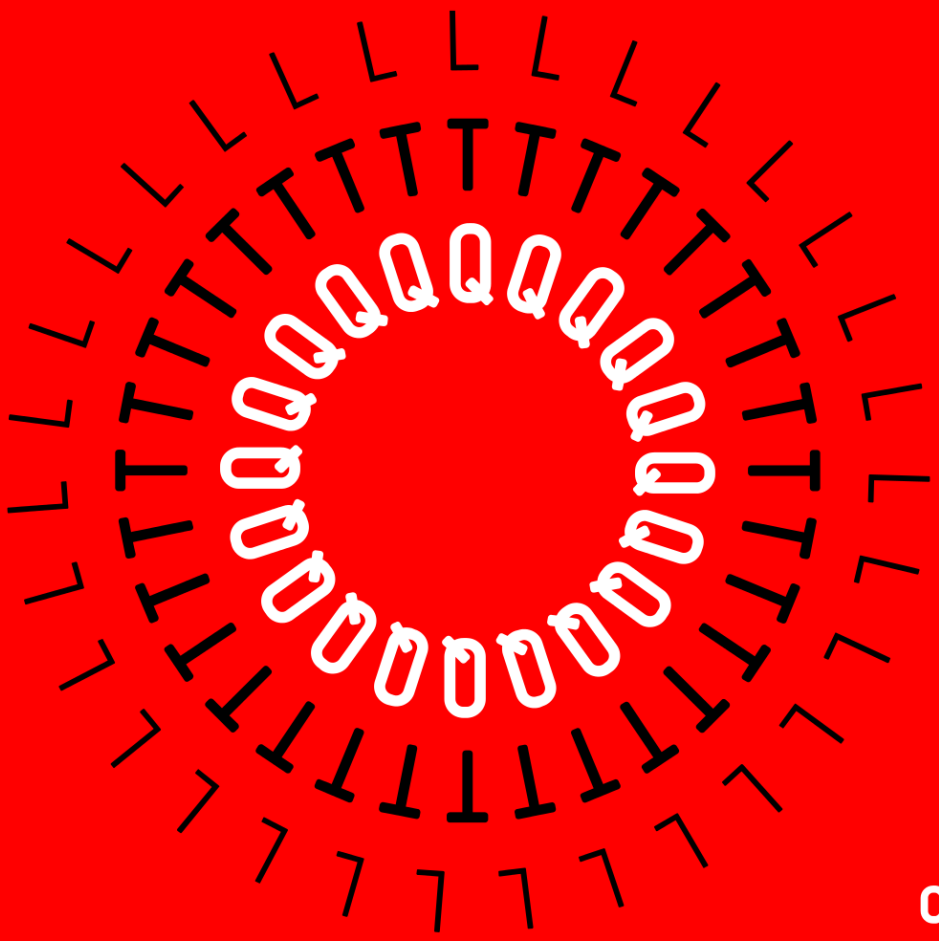


qtleap
quality translation by deep language engineering approaches

Report on the embedding and evaluation of the third MT pilot

DELIVERABLE D3.12

VERSION 2.0 | 2016-10-31

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:

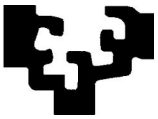
Faculty of Sciences, University of Lisbon

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	Sept. 23, 2016	Rosa Del Gaudio	HF	First draft
	Sept. 26, 2016	Aljoscha Burchardt	DFKI	Various additions
	Sept. 30, 2016	Martin Popel	CUNI	First review and addition of McNemar's test
1.1	Sept. 29, 2016	António Branco	FCUL	Enhancements to the first draft
	Oct. 4, 2016	Aljoscha Burchardt	DFKI	Prefinal touches
	Oct. 6, 2016	António Branco	FCUL	Enhancements to the pre-final draft
	Oct 13, 2016	Eleftherios Avramidis	DFKI	Inter-annotator agreement
2.0	Oct 15-31, 2016	António Branco, Rosa Del Gaudio, Aljoscha Burchardt and Martin Popel	FCUL, HF, DFKI and CUNI	Combining intrinsic and extrinsic evaluation results

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 third MT pilot

DOCUMENT QTLEAP-2016-D3.12
EC FP7 PROJECT #610516

DELIVERABLE D3.12

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, António Branco, Martin Popel

Contents

1 Introduction

Extrinsic evaluation of MT, i.e. assessment of Machine Translation (MT) quality impact within a task other than translation, has not (yet) been established as a major research topic. Reasons may include the prevalent focus of MT research on translation of newspaper texts, which does not readily lend itself to task-based evaluation. In industrial applications of MT, task-based evaluation is certainly performed more frequently, but the results are typically not published.

The evaluation reported in this deliverable joins together general MT research and industrially focused applications of MT, and while covering also intrinsic evaluation, it addresses mostly the extrinsic evaluation of the QTLeap MT Pilot 3, compared with Pilot 0, in the main real-usage scenario addressed in the project (cf. deliverable D3.3).

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 as 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. The final Pilot3, brings deep language processing in support of quality MT, enhanced with deeper semantic information where possible. For four languages, namely Czech, Dutch, Portuguese and Spanish, a fifth MT pilot was developed named Pilot-3-Chimera, in contradistinction to the fourth pilot, named Pilot 3-TectoMT. Chimera is a system combination of TectoMT and Moses. The input text is first translated by TectoMT, thus creating an additional parallel corpus from the input and the output. This is used to construct a secondary phrase table for Moses, which is then applied to the input to produce the translations (Bojar et al. [2013]). More details on the Pilots can be found in project deliverables D2.4, D2.8 and D2.11.

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.

Following the division of labor between WP2 and WP3, intrinsic evaluation was undertaken mostly in WP2 and extrinsic evaluation in WP3.

The intrinsic evaluation of the Pilots 3 against reference translations with automatic measures and using manual error annotation is thoroughly documented in D2.11.

The present deliverable is concerned mostly with the extrinsic (or user) evaluation. It reports also on a part of the intrinsic evaluation performed with human evaluation that was not reported in D2.11. As this manual intrinsic evaluation resorted to the same human evaluators, online forms and experimental apparatus of the extrinsic evaluation, for the sake of the ease of description, it turned out to be more convenient to present it here than in D2.11.

The extrinsic evaluation reported here 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, reported in D3.6, respectively and the experiences gained in the comparative evaluation of subsequent Pilots 1 and 2 (see also Gaudio et al. [2016]). 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. The design of the evaluation has been discussed extensively in the consortium with the goal of finding a good balance

between informativeness of the results and ease of use for the evaluators.

In line with the evaluation of previous MT pilots, the present evaluation includes two distinctive parts. The first part (Section 4) focuses on evaluation how the inbound translation affects the answer retrieval component of the question and answer (QA) algorithm. The second part (Section 5) 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. A third part (Section 6) focuses on assessing the performance of the whole QA system, under the combination of the two previous steps, and the impact of machine translation on that performance.

The two initial sections, in turn, aim at providing contextual information, namely on the embedding of the MT Pilot (Section 2) and on the main objective of the evaluation exercise being reported in this document (Section 3).

The manual 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.

2 Embedding of Pilot 3

The embedding of MT Pilot 3 was performed along the same lines that were followed for the embedding of MT Pilot 0, described in deliverable D3.6, 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

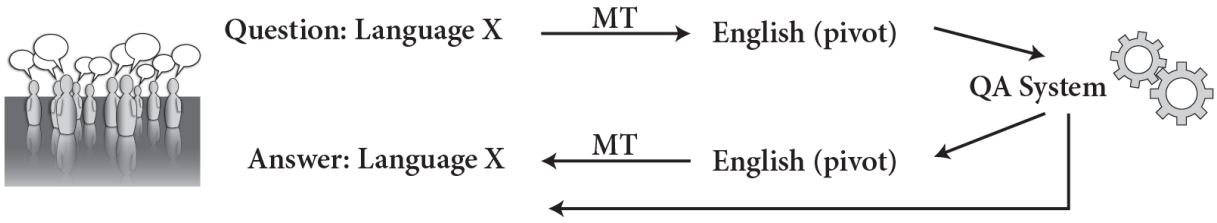


Figure 1: The workflow with the MT services

the list of three 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.

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.

3 The Evaluation's main objective

The main objective of this evaluation is to assess the impact of Pilot3 MT systems in the PcMedic Wizard system. How this impact has been evolving along the project, from Pilot 0 to Pilot 3 is also addressed here.

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 again, this time 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 mostly on each of these steps separately.

4 Extrinsic evaluation: the retrieval step

The main goal of this evaluation step is to compare, in terms of the retrieval performance for stored pairs of questions and answers, 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 into English by the MT services. This evaluation is possible because all the different language versions of questions and answers share a common ID in the database and the algorithm works with English as a pivot language.

As the commercial QA system currently works only with English/Portuguese, there are no data on which to resort to come up with baselines 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 in a given language other than English as it does in the English experiment.

The answers that were delivered for each question were compared 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.

As reported in D3.1 and D3.3, the QTLeap corpus is composed of 4000 question&answer pairs and was collected and human translated 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 fourth batch (called Batch 4). This subcorpus was never used for developing any of the MT pilots in the project, including the last one in the series, that is Pilot 3. As for the evaluation, Batch 4 was used only for this Pilot 3.

All the 1000 questions from this fourth batch were used as input to automatically assess the performance of the retrieval step. The database over which the retrieval procedure was performed included all the 4000 interactions from the entire corpus, in English.

Each question for each language was translated to English using the respective MT Pilot 3 service and the translation was given as input to the QA algorithm. The results obtained was compared with the results obtained when the gold standard English question is used. As this evaluation is automatic and does not need human intervention we compare all the Pilots developed along the project.[1]

4.1 Answering rate per scoring levels

For any input question, the QA system retrieves a list of candidate answers each with a confidence score ranging from 1 to 100, where 100 is taken as indicating that the QA search module is quite sure that the answer is correct for the respective question. The

[1] Chimera version of Pilot 3 was developed only for EN→X translation direction, so it can be evaluated only on the publication step, not the retrieval step being assessed in the present section.

Reference English				
≥95	**75–94**	**≥75**	**50–74**	**25–49**
59.3%	35.8%	**95.1%**	4.5%	0.4%

Table 1: Upper bound of answering rate in retrieval per threshold level: proportion of questions that receive candidate answers in the retrieval step, per level of threshold matched by the questions, when the QA system is fed with the reference 1000 English questions (English being the pivot language to which the questions in other languages are translated)

score represents the confidence level provided by 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 may be low.

Following the PcMedic QA Wizard workflow, and its underlying heuristics and matching procedure, a (previously stored) answer is automatically displayed to the client (without human intervention) only if the input question matches the respective question already stored in the QA repository with a confidence score above 95 points. This is the preferred situation and the major goal of the QA system, as it saves time and money in the long term.

If no stored questions are matched with a confidence score above 95, the top three stored questions with a confidence score above 75 are shown to an operator, who can then choose to adopt one of the respective 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 stored question is matched with a confidence score above 75 points, the input question will be answered by an operator with no help from the QA system. Because of the penalization of old and infrequent questions/answers, it sometimes will happen that a correct stored pair question/answer may receive a score below the thresholds of 95 or of 75.

As the pivot language in the QA system is English and the heuristic is tuned to work with this language, the percentage of answers obtained to our test set in English represents the upper bound of the performance of actual system when dealing with questions in any language other than English. These results are presented in Table 1. The proportion of questions that receive candidate answers in the retrieval step, per level of threshold matched by the questions, are referred to as answering rate in retrieval per threshold level.

The next six tables present the proportion of questions yielding candidate answers per scoring levels, that is the proportion of how many questions the QA algorithm is able to find a candidate answer for within a certain confidence score interval, for each language.

We focus our attention on the first two scores for each language, resumed on the third line (≥75), as it represents a gain in terms of money/time in the real usage scenario.

Starting with Table 2 reporting the results for Basque language, we can see that there is an improvement from Pilot 0 to Pilot 3, and that this improvement applies mostly to the answers scoring above 95. If we consider the scores above 75, Pilot 3 is better then Pilot 0, though this improvement is not constant along the three Pilots, as Pilot 1 presents a lower performance than Pilot 0 and Pilot 2 presents better results regarding the ≥95 answers when compared to the previous two Pilots.

Regarding the Bulgarian language reported in Table 3, Pilot 3 outperforms Pilot 0 on

Basque

Score	P0	P1	P2	P3
≥95	14.3%	14.3%	16.5%	16.6%
75–94	18.4%	16.1%	16.7%	18.0%
≥75	**32.7%**	**30.4%**	**33.2%**	**34.6%**
50–74	47.2%	46.1%	47.6%	47.6%
25–49	19.4%	22.3%	18.8%	16.9%

Table 2: Answering rate in retrieval per threshold level and per Pilot for Basque

Bulgarian

Score	P0	P1	P2	P3
≥95	12.6%	13.8%	12.2%	16.3%
75–94	20.0%	18.3%	16.8%	23.4%
≥75	**32.6%**	**32.1%**	**29.0%**	**39.7%**
50–74	48.8%	50.2%	48.9%	45.6%
25–49	17.6%	16.7%	20.7%	13.7%

Table 3: Answering rate in retrieval per threshold level and per Pilot for Bulgarian

Czech

Score	P0	P1	P2	P3
≥95	15.1%	18.9%	20.9%	20.2%
75–94	22.3%	22.8%	26.4%	25.9%
≥75	**37.4%**	**41.7%**	**47.3%**	**46.1%**
50–74	50.1%	47.3%	44.5%	45.2%
25–49	11.8%	10.7%	7.9%	8.4%

Table 4: Answering rate in retrieval per threshold level and per Pilot for Czech

Dutch

Score	P0	P1	P2	P3
≥95	18.2%	17.2%	16.1%	19.3%
75–94	22.3%	22.4%	21.4%	26.3%
≥75	**40.5%**	**39.6%**	**37.5%**	**45.6%**
50–74	45.8%	47.0%	46.1%	44.0%
25–49	12.4%	13.1%	15.5%	10.3%

Table 5: Answering rate in retrieval per threshold level and per Pilot for Dutch

German

Score	P0	P1	P2	P3
≥95	17.6%	14.3%	–	15.1%
75–94	24.5%	20.1%	–	14.2%
≥75	**42.1%**	**34.4%**	–	**29.3%**
50–74	42.2%	45.8%	–	42.2%
25–49	13.8%	18.0%	–	25.8%

Table 6: Answering rate in retrieval per threshold level and per Pilot for German

both the first two levels of scores. Like for Basque, the path to this improvement is not linear.

In Table 4, for the Czech language, Pilot 2 presents the best results. For this language there is a constant improvement from Pilot 0 to Pilot 1 and from Pilot 1 to Pilot 2, but a slight setback from Pilot 2 to Pilot 3.

For Dutch (Table 5), Pilot 3 outperforms Pilot 0, while Pilot 1 and Pilot 2 both present worst results than Pilot 0.

For German (Table 6), the best results are obtained by Pilot 0. As this language had no resources allocated to it in the Plan of Work of the project (DoW) for WP5, which is the workpackage on lexical semantics whose results helped to leverage Pilot 2 from previous Pilot 1, there was no Pilot 2 for German.

For German Pilot 3, the inbound direction addressed here is a special case. Towards the end of QTLeap, neural MT had become a hot topic driven on the one hand by announcements of large companies and at the same time by good performance of academic NMT systems, e.g. in WMT 2016. The QTLeap project therefore decided to dedicate a few person-months for the experiment of setting up an NMT system trained on the same data as the German Pilot 0 (see D2.11 for a description of the system). This has been done as we were interested in assessing to what extent today's neural networks technology could compare with Moses and our deep MT systems in this domain-specific cross-lingual information retrieval task.

For uniformity, we called this system Pilot 3, even if it follows a totally different approach from the remainder of the languages in the project other than German.

Portuguese

Score	P0	P1	P2	P3
≥95	13.4%	10.7%	15.9%	16.4%
75–94	17.9%	15.1%	20.7%	18.7%
≥75	**31.3%**	**25.8%**	**36.6%**	**35.1%**
50–74	47.9%	51.3%	47.8%	48.0%
25–49	18.8%	20.9%	14.8%	16.2%

Table 7: Answering rate in retrieval per threshold level and per Pilot for Portuguese

In Table 7, Portuguese Pilot 3 outperforms all the other three Pilots at the ≥95 threshold. From Pilot 0 to Pilot 1 there is a worsening, but after that the improvement is constant along the subsequent Pilots.

Spanish

Score	P0	P1	P2	P3
≥95	16.2%	11.8%	19.0%	18.8%
75–94	22.6%	17.0%	21.9%	22.1%
≥75	**38.8%**	**28.8%**	**40.9%**	**40.9%**
50–74	46.8%	48.7%	44.6%	44.4%
25–49	13.2%	20.5%	13.2%	13.4%

Table 8: Answering rate in retrieval per threshold level and per Pilot for Spanish

Regarding the Spanish language reported in Table 8, there is an overall improvement of Pilot 3 over Pilot 0, due to the improvement of answers scoring ≥95. Furthermore, for

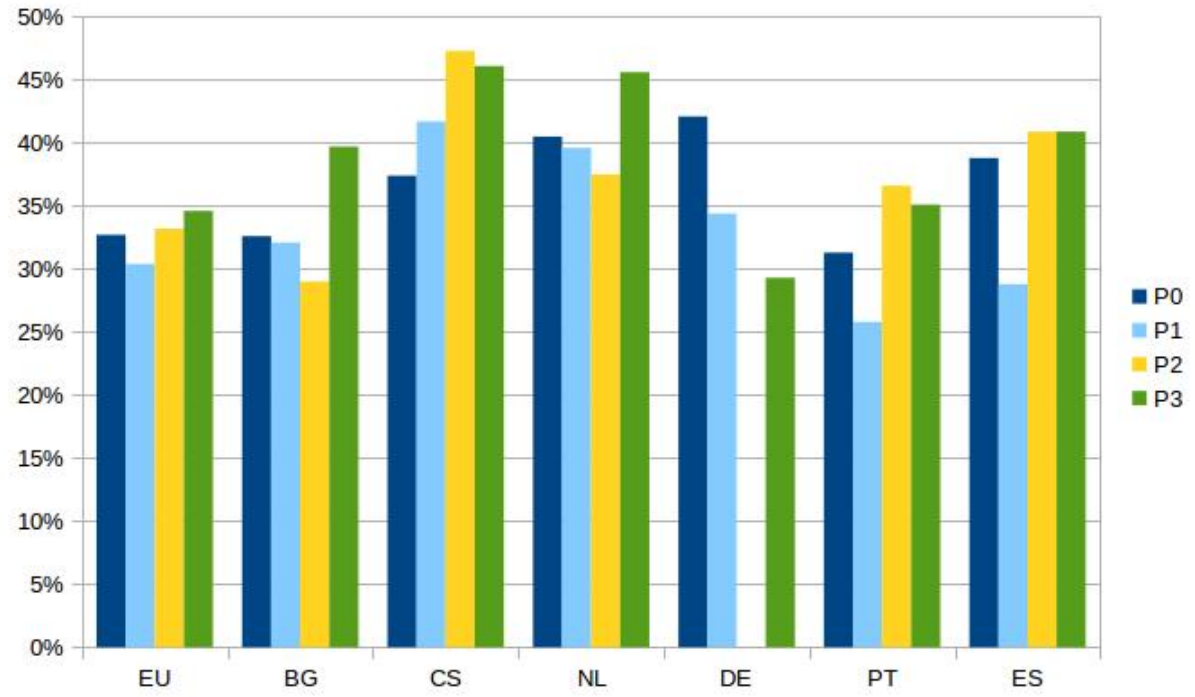


Figure 2: Answering rate in retrieval per language and per Pilot when the threshold of questions is ≥75

this language, Pilot 1 scores below Pilot 0 and the improvements over Pilot 0 appear with Pilot 2.

Figure 2 summarizes all the results of the answering rate in the retrieval step for questions scoring above 75. In general, Pilot 3 outperforms Pilot 0 for almost all languages, with the exception of German.

Pilot 3 also outperforms the other two Pilots 1 and 2, with the only exception of Czech, whose Pilot 2 presents better results than Pilot 3, and Spanish, whose Pilot 2 and Pilot 3 present similar scores.

If we consider the average performance of a system addressing all the languages, but German, we get an average of 35.6% for Pilot 0, 33.1% for Pilot 1, 37.4 for Pilot 2 and 40.3% for Pilot 3. This means that from Pilot 0 to Pilot 3, there was an improvement of almost 5 percentage points.

We left results for German out of the average due to the specific choices made in the development of the German MT system (as explained before) that makes the German Pilot 3 not comparable with the other ones.

4.2 Accuracy per ranking positions

While the scores in the rows for the ≥95 threshold in the Tables of the previous Section 4.1 indicate the proportion of input questions for which the QA system automatically delivers an answer, these scores do not provide an indication of the quality of this automatic answer in relation to the gold standard, i.e. whether the high-confidence automatically retrieved question&answer pair contains the correct answer for the input question.

Tables 9 to 15 report, for each language other than English and per Pilot, the proportion of input questions to which the correct, gold standard answer appears in the list of

Basque

Score	P0	P1	P2	P3
First (%)	49.6%	38.0%	43.2%	45.9%
Average score	71.5	70.4	71.2	71.6
First 2 (%)	60.9%	50.7%	55.6%	58.5%
Average score	68.3	65.3	67.5	68.0
First 3 (%)	67.1%	57.6%	62.8%	66.2%
Average score	66.7	63.4	65.5	66.1

Table 9: Accuracy in retrieval per ranking positions and per Pilot for Basque

Bulgarian

Score	P0	P1	P2	P3
First (%)	55.7%	51.8%	49.2%	57.9%
Average score	68.8	68.9	66.3	73.0
First 2 (%)	66.3%	63.9%	60.2%	70.2%
Average score	66.3	66.2	63.9	70.0
First 3 (%)	72.3%	70.9%	67.6%	76.5%
Average score	65.0	64.8	62.4	68.4

Table 10: Accuracy in retrieval per ranking positions and per Pilot for Bulgarian

the top three answers automatically retrieved without using any threshold. Hence, given the ranked list of pairs of questions/answers automatically retrieved from the repository, the tables present information on the probability of the gold standard answer appearing in the first, in the first two or in the first three positions. To this proportion, we refer as accuracy in retrieval per ranking positions, and briefly designate it with the acronym AccuracyPRP. And to designated the accuracy in retrieval in the first n ranked positions, we use the acronym AccuracyPRP@n.

The tables also show the average score of the correct answers for each case. In the second line of the tables, for example, the figure shows the average score for all correct answers that are retrieved in the first position.

English scores are constant, because English represents the gold standard and thus it always represents the first best answer and its score. For this reason, the accuracy is always 100% and the score is 94.2, and represents the upper bound.

Czech

Score	P0	P1	P2	P3
First (%)	61.3%	63.3%	68.1%	67.5%
Average score	71.9	73.9	76.6	76.0
First 2 (%)	73.9%	74.6%	79.5%	79.0%
Average score	69.8	71.7	74.3	73.6
First 3 (%)	79.0%	80.6%	84.0%	85.0%
Average score	68.4	70.2	73.3	72.2

Table 11: Accuracy in retrieval per ranking positions and per Pilot for Czech

Dutch				
Score	**P0**	**P1**	**P2**	**P3**
First (%)	61.5%	60.9%	59.2%	65.0%
Average score	73.8	72.0	70.4	74.5
First 2 (%)	74.3%	74.9%	72.6%	79.0%
Average score	71.0	69.2	68.0	71.7
First 3 (%)	80.0%	79.6%	77.8%	83.8%
Average score	69.6	68.2	66.5	70.5

Table 12: Accuracy in retrieval per ranking positions and per Pilot for Dutch

German				
Score	**P0**	**P1**	**P2**	**P3**
First (%)	67.3%	64.1%	–	40.1%
Average score	73.8	69.9	–	66.8
First 2 (%)	79.2%	75.9%	–	54.8%
Average score	71.3	67.2	–	62.9
First 3 (%)	83.3%	80.1%	–	62.7%
Average score	70.5	66.0	–	60.9

Table 13: Accuracy in retrieval per ranking positions and per Pilot for German

Portuguese				
Score	**P0**	**P1**	**P2**	**P3**
First (%)	51.0%	50.8%	57.0%	52.9%
Average score	68.0451	66.5	71.1	71.1
First 2 (%)	63.9%	62.4%	70.8%	64.7%
Average score	65.1189	63.8	67.8	68.1
First 3 (%)	70.2%	69.3%	76.6%	71.8%
Average score	64.1866	62.4	67.0	66.5

Table 14: Accuracy in retrieval per ranking positions and per Pilot for Portuguese

Spanish				
Score	**P0**	**P1**	**P2**	**P3**
First (%)	60.9%	54.8%	59.5%	59.9%
Average score	71.8	68.0	74.7	74.3
First 2 (%)	74.9%	65.9%	72.7%	73.3%
Average score	69.5	66.0	71.8	71.5
First 3 (%)	79.9%	72.0%	77.9%	78.3%
Average score	68.5	64.3	70.2	70.0

Table 15: Accuracy in retrieval per ranking positions and per Pilot for Spanish

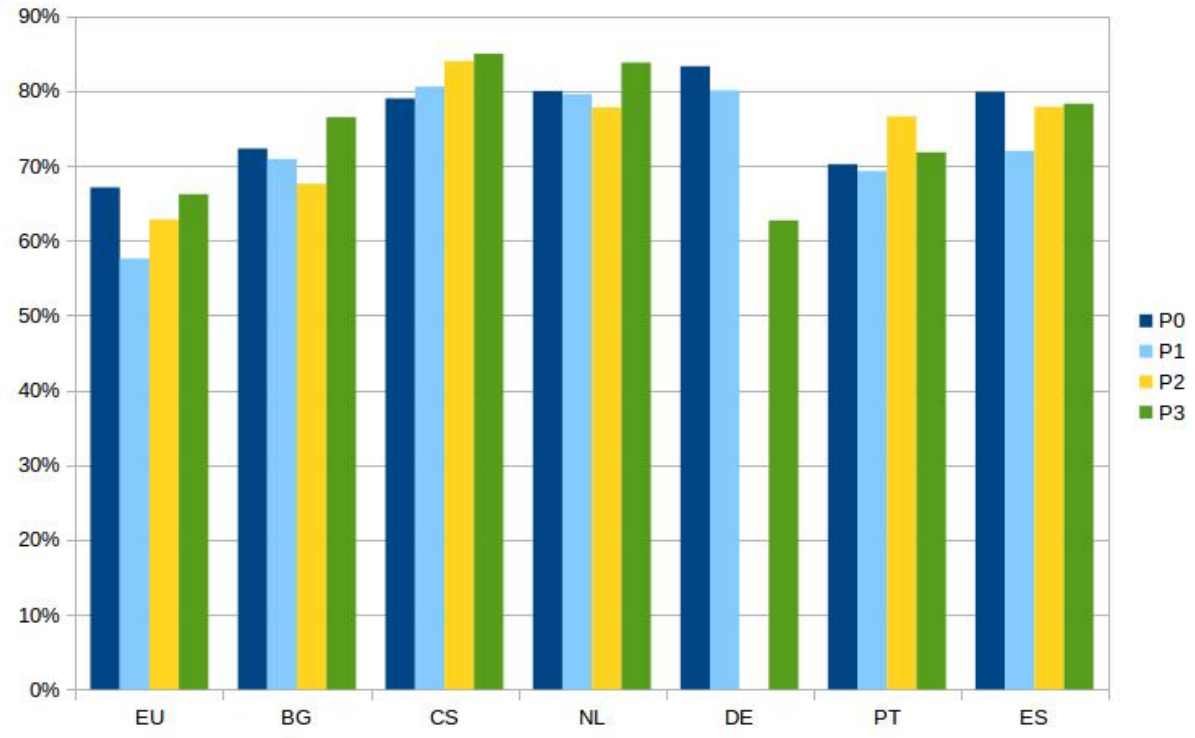


Figure 3: Accuracy in retrieval of answers appearing in the first three positions (AccuracyPRP@3)

For the Basque (Table 9) and Spanish (Table 15) languages, Pilot 0 outperforms all the other Pilots. From Pilot 0 to Pilot 1 there is a worsening, and even if Pilot 2 and Pilot 3 present constant improvements, they are not enough to outperform Pilot 0.

Both Bulgarian (Table 10) and Dutch (Table 12) Pilots 3 present better results than all the other Pilots. However, Pilot 1 or Pilot 2 do not improve over Pilot 0.

In Table 11, similarly to what happens with answering rate, Czech Pilot 3 outperforms Pilot 0 and Pilot 1, but not Pilot 2, which presents slightly better results in the first-position accuracy. In the first-three-positions accuracy, Pilot 3 obtains the best results.

Regarding the German language, reported in Table 13, Pilot 0 presents better results than both Pilot 1 and Pilot 3. The neural MT system (Pilot 3 here) performs considerably worse than Moses (Pilot 0) and than the deeper MT system (Pilot 1).

In Table 14, Portuguese Pilot 3 outperforms Pilot 0 and Pilot 1, but not Pilot 2.

Figure 3 presents a graphical summary of the accuracy per ranking position for all the languages. In general, Pilot 3 outperforms Pilot 0 in terms of accuracy in the retrieval step for almost all languages, with the exception of Basque, German and Spanish. Pilot 3 also outperforms the other two Pilots 1 and 2, with the exception of Portuguese, whose Pilot 2 present slightly better results than Pilot 3.

If we consider the average performance of a system addressing all the languages, but German, we get an average of 74.8% for Pilot 0, 71.7% for Pilot 1, 74.5% for Pilot 2 and 76.9% for Pilot 3. This means that from Pilot 0 to Pilot 3, there was an improvement of more than 2 percentage points in average.

4.3 Accuracy per scoring levels

In the previous section, the values for accuracy per ranking positions were reported. It is important to determine the accuracy for the different thresholds, that is the proportion of answers that are retrieved and correct in the first ranked position and whose score is 95 or more, is between 75 and 94 or is below 75. To this proportion, we refer as accuracy per scoring level, and use the acronym AccuracyPSL. The acronym AccuracyPSL≥n is used to refer to the accuracy for answers scoring n or more.

Table 16 shows the results for accuracy per scoring level when the reference English is used. Given the English pairs of questions and respective answers form the reference repository of the QA system, these figures are identical to the figures in Table 1 concerning answering rate per threshold levels.

Reference English

≥95	**75–94**	**≥75**	**<75**
59.3%	35.8%	**95.1%**	4.9%

Table 16: Accuracy in retrieval per scoring levels for English

Tables 17 to 23 report on the accuracy per scoring level for the other languages in the project.

For Basque (Table 17), Pilot 3 presents better results for answers whose retrieval score is ≥95 when compared with Pilot 0, but the results are worse for answers scoring between 75 and 94. When these two scores are combined, Pilot 0 slightly outperforms Pilot 3.

For Bulgarian (Table 18), Dutch (Table 20) and Spanish (Table 23), Pilot 3 outperforms all the other Pilots on both scoring levels.

For Czech (Table 19) and Portuguese (Table 22), Pilot 2 gets better results than the other Pilots, followed by Pilot 3.

For German, (Table 21), Pilot 0 obtains by large the best results.

Basque

Score	**P0**	**P1**	**P2**	**P3**
≥95	8.7%	6.3%	8.5%	9.4%
75–94	10.9%	7.8%	8.9%	9.6%
≥75	**19.6%**	**14.1%**	**17.4%**	**19.0%**
<75	80.4%	85.9%	82.6%	81.0%

Table 17: Accuracy in retrieval per scoring level and per Pilot for Basque

Bulgarian

Score	P0	P1	P2	P3
≥95	8.4%	8.1%	6.3%	10.8%
75–94	12.6%	11.5%	9.4%	16.6%
≥75	**21.0%**	**19.6%**	**15.7%**	**27.4%**
<75	79.0%	80.4%	84.3%	72.6%

Table 18: Accuracy in retrieval per scoring level and per Pilot for Bulgarian

Czech

Score	P0	P1	P2	P3
≥95	9.5%	12.6%	15.7%	14.9%
75–94	15.8%	16.0%	19.9%	19.7%
≥75	**25.3%**	**28.6%**	**35.6%**	**34.6%**
<75	74.7%	71.4%	64.4%	65.4%

Table 19: Accuracy in retrieval per scoring level and per Pilot for Czech

Dutch

Score	P0	P1	P2	P3
≥95	13.4%	11.4%	10.0%	13.8%
75–94	15.7%	15.2%	14.0%	19.0%
≥75	**29.1%**	**26.6%**	**24.0%**	**32.8%**
<75	70.9%	73.4%	76.0%	67.2%

Table 20: Accuracy in retrieval per scoring level and per Pilot for Dutch

German

Score	P0	P1	P2	P3
≥95	13.3%	10.0%	-	5.5%
75–94	19.7%	16.0%	-	6.0%
≥75	**33.0%**	**26.0%**	**-**	**11.5%**
<75	67.0%	74.0%	-	88.5%

Table 21: Accuracy in retrieval per scoring level and per Pilot for German

Portuguese

Score	P0	P1	P2	P3
≥95	6.1%	5.6%	8.8%	8.7%
75–94	11.3%	10.4%	15.2%	13.6%
≥75	**17.4%**	**16.0%**	**24.0%**	**22.3%**
<75	82.6%	84.0%	76.0%	77.7%

Table 22: Accuracy in retrieval per scoring level and per Pilot for Portuguese

Spanish

Score	P0	P1	P2	P3
≥95	8.4%	8.1%	6.3%	10.8%
75–94	12.6%	11.5%	9.4%	16.6%
≥75	**21.0%**	**19.6%**	**15.7%**	**27.4%**
<75	79.0%	80.4%	84.3%	72.6%

Table 23: Accuracy in retrieval per scoring level and per Pilot for Spanish

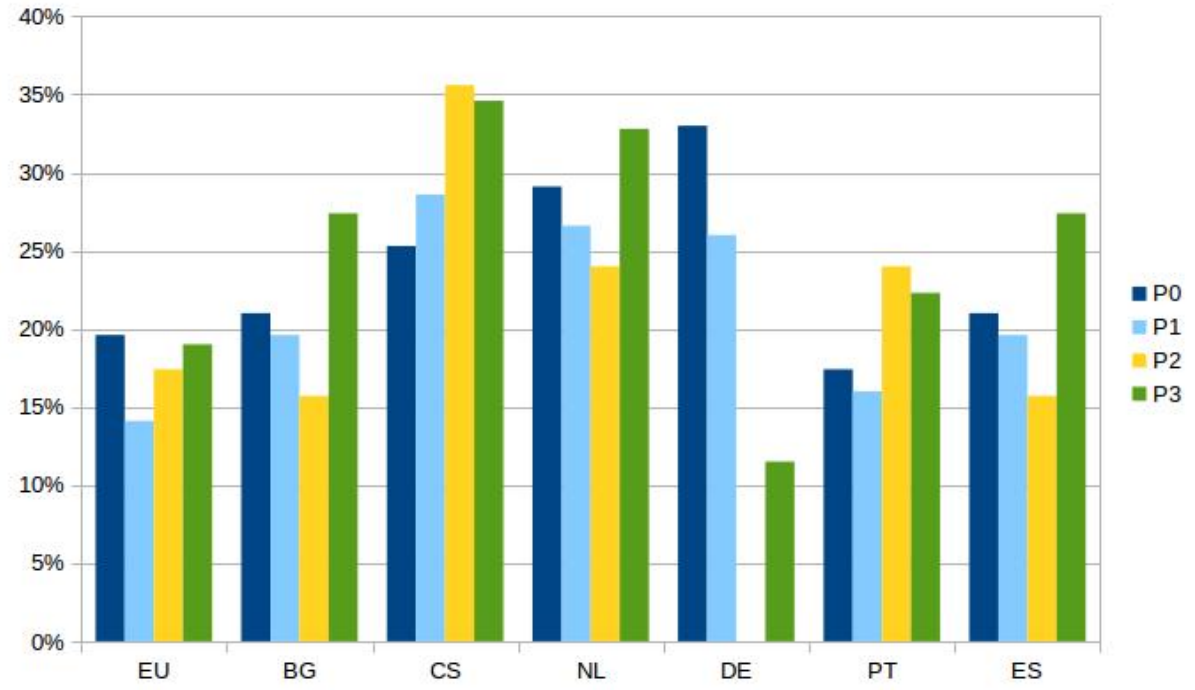


Figure 4: Accuracy in retrieval for the scoring level ≥75 (AccuracyPSL≥75) per Pilot and per language

Figure 4 shows a graphical summary of the accuracy for questions scoring ≥75. Considering the average, there is an evolution from 22.2%, for Pilot 0, to 20.8%, for Pilot 1, to 22.1%, for Pilot 2, and finally to 27.3%, for Pilot 3, when the average accuracy ≥75 is taken into account, excluding German for the reasons explained above, related to the fact that Pilot 3 for German resulted for the experimentation with neural machine translation technology.

This means that from Pilot 0 to Pilot 3, there was an improvement of more than 5 percentage points in average in terms of accuracy in retrieval for the scoring level ≥75.

4.4 Comparison with intrinsic evaluation

In this Section, we compare the results obtained with the retrieval step with the results of the intrinsic evaluation of the inbound (X→EN) translation direction.

The intrinsic evaluation in D2.11 reports on assessing translation quality of the MT Pilots with several automatic measures and also some human evaluation. While the extrinsic evaluation and the intrinsic evaluations serve different purposes and are complementary by design, the results have been compared where possible.

In Table 24 and the companion Figure 5, the results of the extrinsic evaluation are presented together with the intrinsic BLEU scores of Pilot 0 and Pilot 3 (taken from D2.11). As representative metrics of the extrinsic evaluation exercise, we select the proportion of questions yielding candidate answers with a score of ≥75 (*AnsweringRate≥75*), the proportion of correct answers in the first three positions (*AccuracyPRP@3*) and the proportion of questions with the correct answer in first ranked position of yielded candidate answers with scores ≥75 (*AccuracyPSL ≥75*).

	EU	**BG**	**CS**	**NL**	**DE**	**PT**	**ES**
AnsweringRate(≥75) P0	32.7%	32.6%	37.4%	40.5%	42.1%	31.3%	38.8%
AnsweringRate(≥75) P3	34.6%	39.7%	46.1%	45.6%	29.3%	35.1%	40.9%
AccuracyPRP (@3) P0	67.1%	72.3%	79.0%	80.0%	83.3%	70.2%	79.9%
AccuracyPRP (@3) P3	66.2%	76.5%	85.0%	83.8%	62.7%	71.8%	78.3%
AccuracyPSL (≥75) P0	19.6%	21.0%	25.3%	29.1%	33.0%	17.4%	25.9%
AccuracyPSL (≥75) P3	19.0%	27.4%	34.6%	32.8%	11.5%	22.3%	28.0%
BLEU P0	13.70	18.54	20.53	27.89	34.74	13.75	26.88
BLEU P3	7.30	24.93	21.31	30.34	24.51	9.72	18.07
Δ AnsweringRate (≥75)	1.9	7.1	8.7	5.1	−12.8	3.8	2.1
Δ AccuracyPRP (@3)	−0.9	4.2	6.0	3.8	−20.6	1.6	−1.6
Δ AccuracyPSL (≥75)	−0.6	6.4	9.3	3.7	−21.5	4.9	2.1
Δ BLEU	−6.40	6.39	0.78	2.45	−10.23	−4.03	−8.81

Table 24: Comparison of extrinsic and intrinsic evaluation results in the retrieval step (table): Comparison of the differences between Pilot 0 and Pilot 3 in terms of the results from intrinsic evaluation of the translation direction (X→EN) and of the results from extrinsic evaluation of the retrieval step, per language

In order to compare the results of Pilot 0 and Pilot 3, the difference between those results (Δ) for the four different metrics was computed. Figure 5 shows the graphical representation of these Δ figures.

For most languages, the intrinsic and extrinsic results show Δ of identical signal. For Bulgarian, Czech and Dutch all the metrics agree in indicating that Pilots 3 and the QA systems supported by them are performing better than QA systems supported by Pilot 0. In the case of Czech, there is a more substantial gain in retrieval performance given the improvement obtained in terms of machine translation than in the other two languages mentioned above.

For Basque, the signal of ΔBLEU is identical to the signal of ΔAccuracyPRP@3 and ΔAccuracyPSL≥75, while for Spanish, that similarity obtains with ΔAccuracyPRP@3 only.

For Portuguese, the extrinsic Δs and the intrinsic Δ point towards opposite directions.

Regarding German, both ΔBLEU and the extrinsic Δs seem to confirm the general

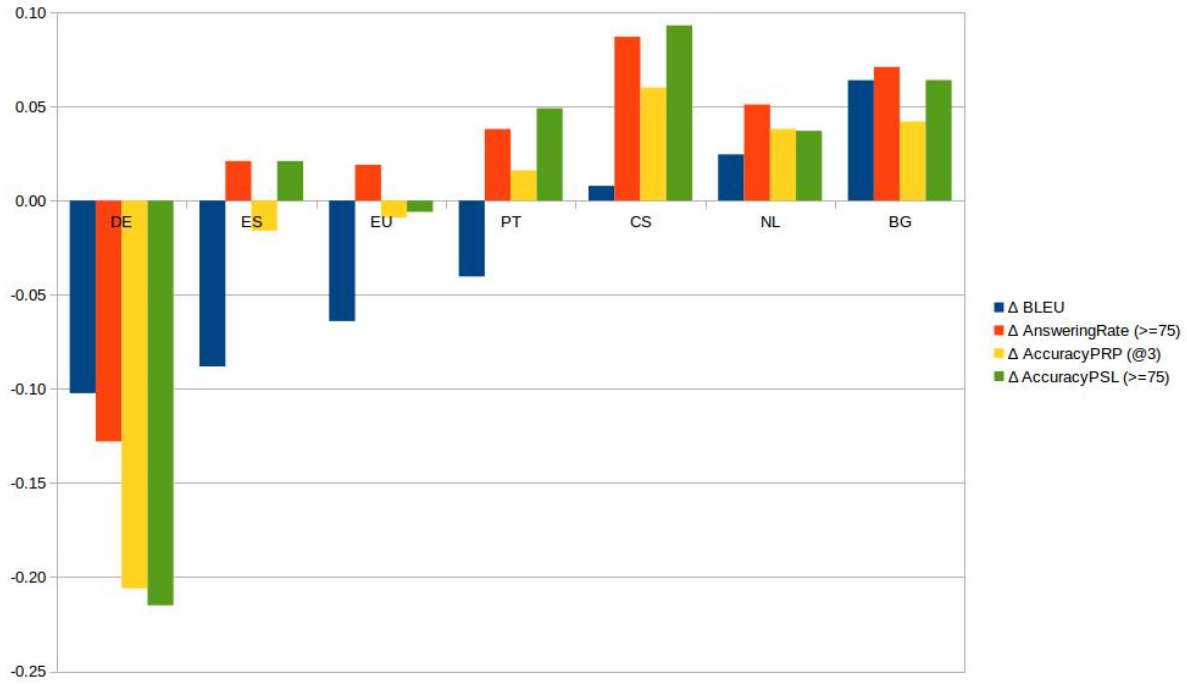


Figure 5: Comparison of extrinsic and intrinsic evaluation results in the retrieval step (synopsis in chart): Comparison of the differences between Pilot 0 and Pilot 3 in terms of the results from intrinsic evaluation of the translation direction (X→EN) and of the results from extrinsic evaluation of the retrieval step, per language

observation that current neural MT does not easily cope with domain data that well (which is typically much smaller in proportion to the general domain training data).

It is difficult to proceed with more detailed analysis of these results. On the one hand, there is no baseline (HF company did not offer multilingual services in the past) and on the other hand, as the matching algorithm is business secret, one can only speculate about the reasons for increased or decreased values across pilots for the different languages.

As a final remark in thus subsection, it is worth note again that the priority of the project was to improve outbound translation quality into languages other than English, rather than the inbound direction, whose results are presented above.

5 Extrinsic evaluation: the publication step

In the design of the evaluation of the outbound, publication step multiple issues had to be taken in consideration.

First, as the PcMedic Wizard QA application is currently used for the Portuguese/English, no real baseline to evaluate the business case in a wider multilingual scenario was available.

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

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

The evaluation set-up of the publication step for Pilot 3 follows closely the evaluation

of Pilot 0, based on the metrics presented in deliverable D3.3.

In the present evaluation exercise, the automatically translated answers produced by Pilot 0 and Pilot 3 are first rated in isolation by the experimental subjects recruited. We decided to repeat the evaluation of Pilot 0 to control for difference in this evaluators cohort and also to account for the fact that a "fresh" batch of the QTLeap corpus was used in this evaluation exercise.

In order to get direct comparative information about the relation between the two Pilots under consideration, evaluators were subsequently asked to compare both pilots. This type of comparative evaluation had been performed in previous evaluations of Pilot 1 and Pilot 2, respectively (see D3.8 and D3.10).

All project partners helped in this evaluation by translating the online 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 IT experts.

5.1 Experimental setup

5.1.1 Evaluation interface

A new web-based interface was set up and translated for all the languages. This interface for the extrinsic evaluation of Pilot 3 is available at `http://83.240.145.199/questionnaire/pilot3/`. On the first page (Figure 6), evaluators select their language.

Figure 6: Evaluation Interface: Selecting a Language

The second page (Figure 7) provides a brief explanation of the evaluation task. The evaluators provide their e-mail addresses 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 is asked for basic information about age, sex, education, and familiarity with information technology (Figure 8).

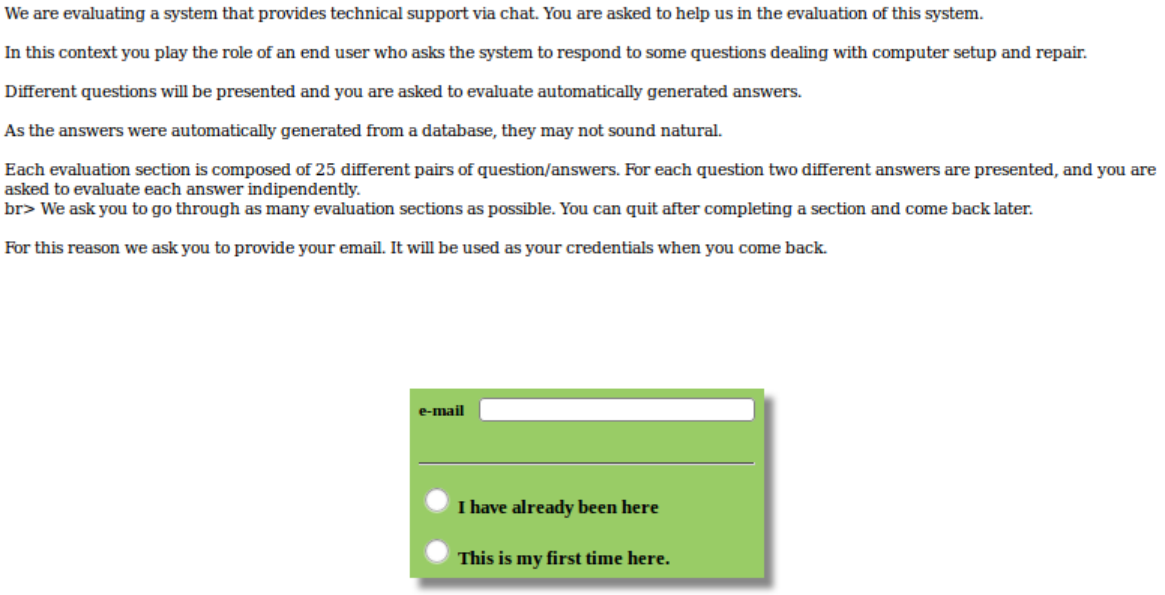


Figure 7: Evaluation Interface: Introduction and log in

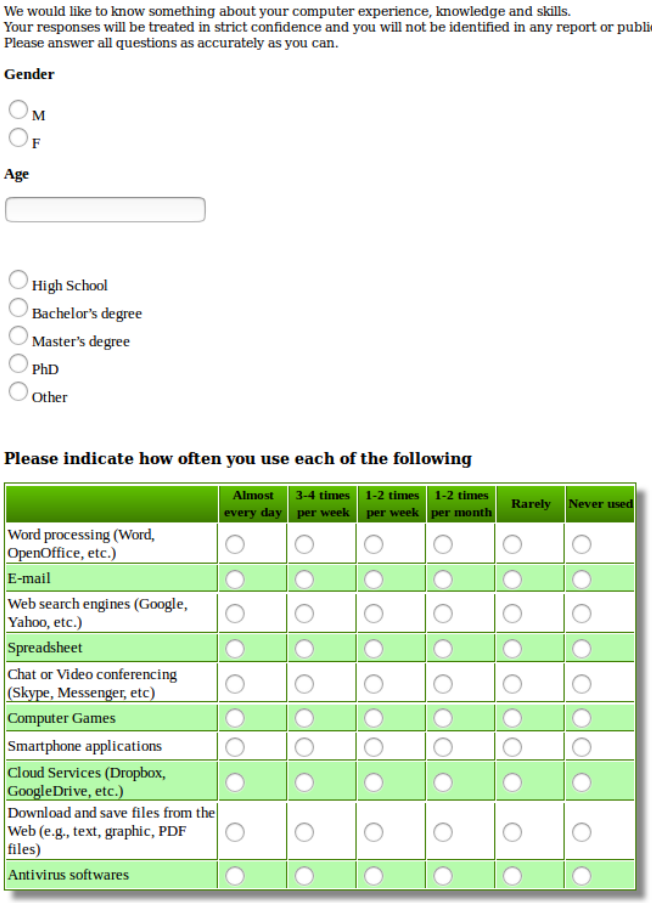


Figure 8: Evaluation Interface: Basic information

The next page presents Form 1 (Figure 9), and represents the start of the evaluation. Here a question from the QTLeap corpus is presented in the selected language and the evaluator is asked to provide a self-estimation of his knowledge level (high, medium, low) about the subject involved in the question.

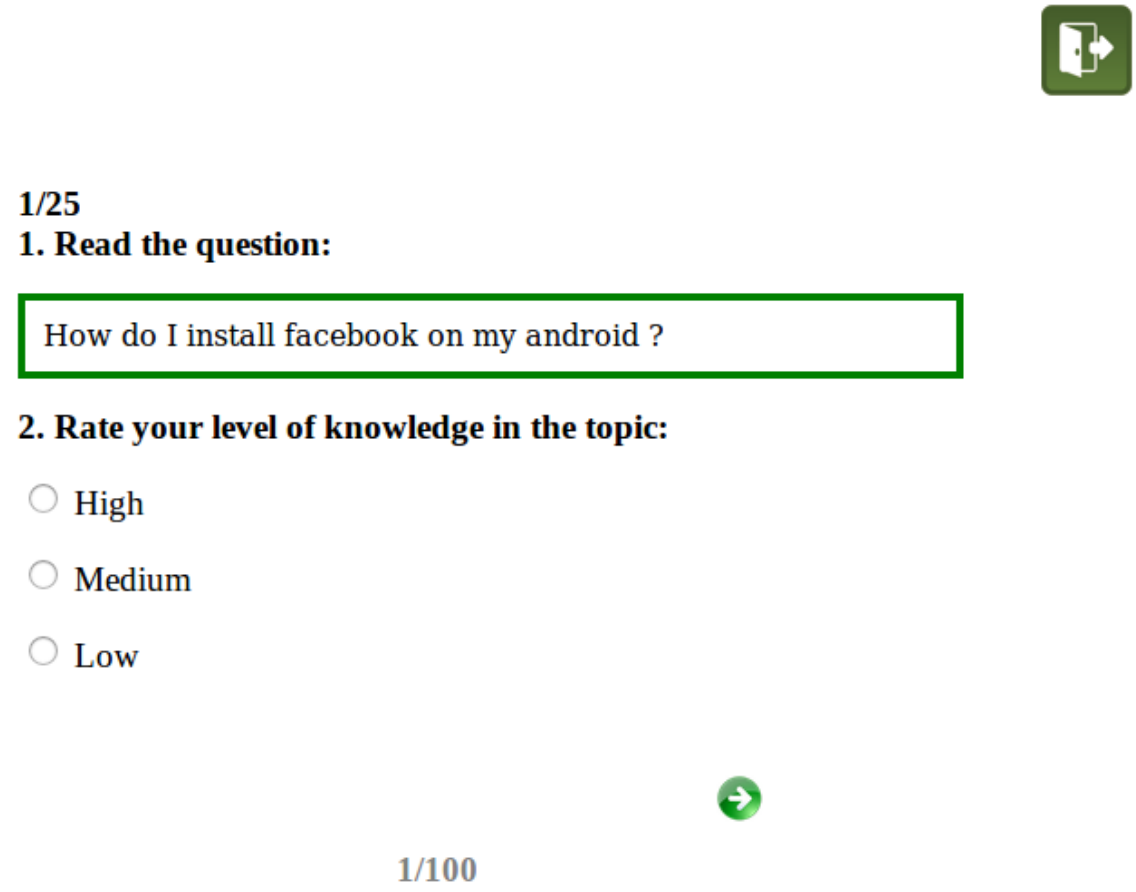


Figure 9: Evaluation Interface: Form 1

In Form 2 (Figure 10), the two automatically translated answers (A and B) generated by Pilot 0 and Pilot 3 are presented (in randomized order). The evaluator is asked to assess the usefulness of answers A and B independently, selecting one of the following options for each answer:

- It would clearly help me solve my problem / answer my question
- It might help, but would require some thinking to understand it.
- Is not helpful / I don't understand it

In form 3 (Figure 11), the question and the (automatically translated) answers are shown again, followed by the reference answer. This time the subject is asked to compare the manually translated, gold standard answer with each automatic answer. The subject is asked to re-evaluate the answers A and B selecting one of the following options:

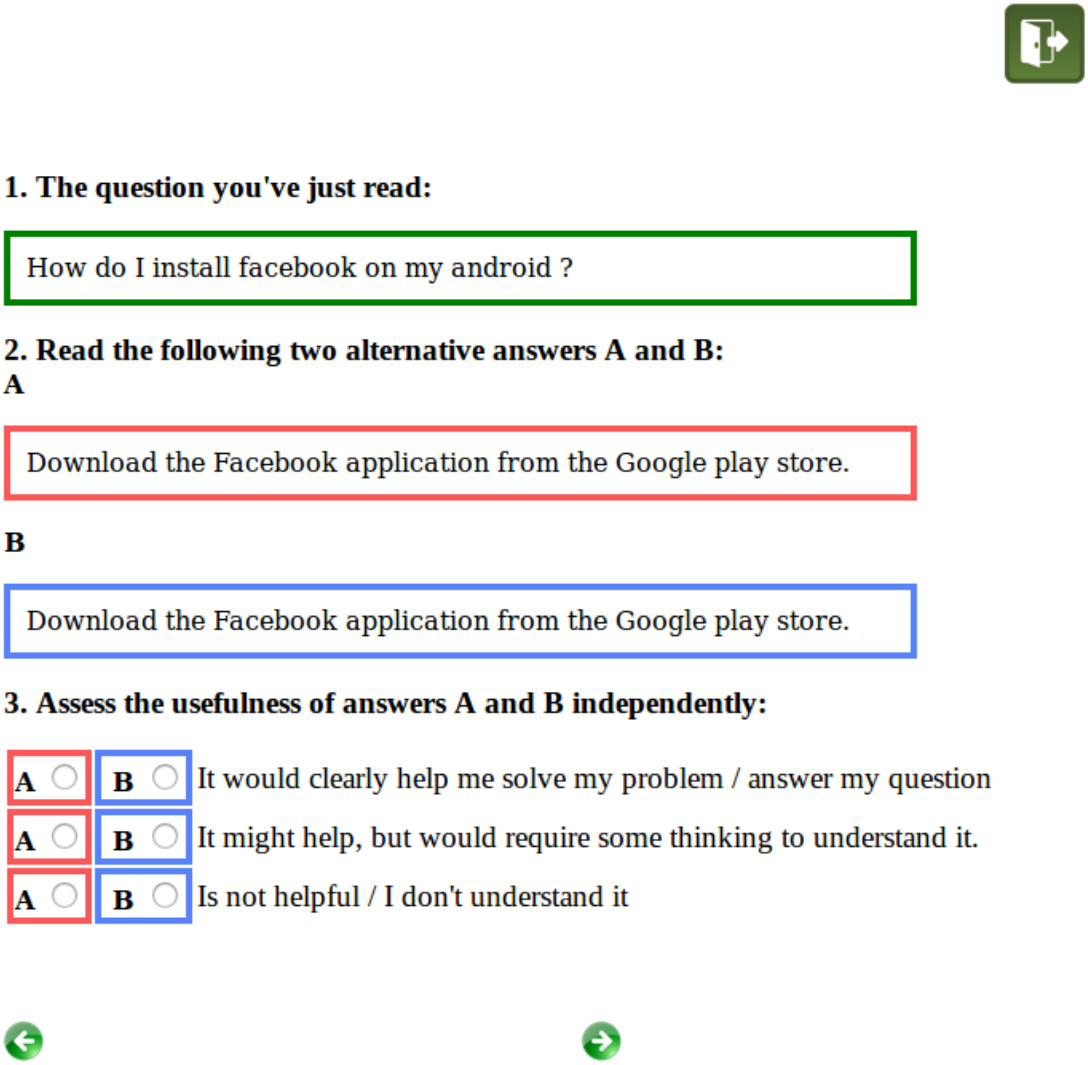


Figure 10: Example of evaluation Interface: Form 2

- gives the right advice.
- gets minor points wrong.
- gets important points wrong.

Finally in form 4 (Figure 12), evaluators are asked to compare answers A and B, selecting one of the following options:

- A is a better answer than B
- B is a better answer than A
- A and B are equally good answers
- A and B are equally bad answers

1. The question:

How do I install facebook on my android ?

2. The two alternative answers A and B you have alredy read:
A

Download the Facebook application from the Google play store.

B

Download the Facebook application from the Google play store.

3. Now read the reference answer C:

Download the Facebook application from the Google play store.

4. Considering answer C as giving the correct information, which of the following is true about the answers A and :

A ○ **B** ○ gives the right advice.
A ○ **B** ○ gets minor points wrong.
A ○ **B** ○ gets important points wrong.


Figure 11: Example of evaluation Interface: Form 3

1. The question:

How do I install facebook on my android ?

2. The reference answer C:

Download the Facebook application from the Google play store.

3. The two alternative answers A and B you have alredy read:
A

Download the Facebook application from the Google play store.

B

Download the Facebook application from the Google play store.

4. Suppose that the reference answer is correct. Does **A** or **B** provide a better answer to the question?

- **A** is a better answer than **B**
- **B**is a better answer than **A**
- **A** and **B** are equally good answers
- **A** and **B** are equally bad answers


Figure 12: Example of evaluation Interface: Form 4

EU	BG	CS	NL	DE	PT	ES
3.0	3.0	2.2	2.9	2.0	2.1	2.0

Table 25: The average number of evaluations per interaction per language

Table 26 shows the basic demographic information about the pool of evaluators.

Sex		Age	Instruction				
F	M		High School	Bachelor	Master	Phd	Other
55.9%	44.1%	35.9	23.5%	23.5%	38.2%	5.9%	8.8%

Table 26: Information about the profile of volunteer evaluators

5.1.2 Evaluated content and evaluators

This evaluation was carried out in a controlled setting in order to avoid having to deal with different experimental variables that might 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 the QTLeap Corpus[2] were used.[3] Each project partner recruited volunteers that were not IT experts in order to simulate the typical user of the PcMedic Wizard QA 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 2.5 evaluations per interaction. Table 25 presents the average of evaluations for each language

Table 26 displays information on the profile of the evaluators in terms of sex, age and education.

Table 27, in turn, shows data on the use of technology by evaluators. This gives us another characterization of the population of evaluators. In general, people participating in this evaluation use a computer or a smarthphone almost every day. They use common applications such as e-mail, web-engines and word processors. When we look at more specific and/or advanced applications, such as spreadsheet or cloud services, the use goes down. This indicates that even if the use of the technology is carried out on a daily basis, we are in presence of laypersons, that use computers for simple tasks.

[2] `http://metashare.metanet4u.eu/go2/qtleapcorpus`

[3] In the project, we have divided the corpus into four batches of 1000 interactions. The evaluation used interactions from Batch 4.

	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.)	56.7%	16.7%	13.3%	10.0%	0.0%	3.3%
E-mail	90.0%	6.7%	0.0%	3.3%	0.0%	0.0%
Web search engines (Google, Yahoo, etc.)	86.7%	3.3%	10.0%	0.0%	0.0%	0.0%
Spreadsheet	16.7%	23.3%	13.3%	10.0%	26.7%	10.0%
Chat or Video conferencing (Skype, Messenger, etc)	33.3%	10.0%	16.7%	20.0%	13.3%	6.7%
Computer Games	16.7%	16.7%	3.3%	3.3%	36.7%	23.3%
Smartphone applications	60.0%	6.7%	16.7%	3.3%	3.3%	10.0%
Cloud Services (Dropbox, GoogleDrive, etc.)	23.3%	10.0%	26.7%	16.7%	6.7%	16.7%
Download and save files from the Web	46.7%	30.0%	10.0%	3.3%	6.7%	3.3%
Antivirus software	33.3%	10.0%	6.7%	20.0%	23.3%	6.7%

Table 27: Characterization of evaluators on the use of technology

5.2 Results and Discussion

This section presents the results of the extrinsic evaluation in the publication step.

Czech, Dutch, Portuguese and Spanish languages present two sets of evaluations. We carried on two separate evaluations for these languages, one using Pilot3-TectoMT and the other using Pilot3-Chimera.

Table 28 reports on the average of the self-estimation of the knowledge (high, medium, low) of the subjects on the different questions, per language.

Knowledge	**EU**	**BG**	**CS**	**NL**	**DE**	**PT**	**ES**
High	22.0%	42.4%	0.3%	19.9%	25.6%	3.5%	52.4%
Medium	33.4%	31.1%	7.1%	49.1%	21.2%	36.3%	22.8%
Low	44.6%	26.5%	92.6%	31.0%	53.2%	60.2%	24.8%
Medium+Low	78.0%	57.6%	99.7%	80.1%	74.4%	96.5%	47.6%

Table 28: The average self-estimated level of the knowledge by the evaluators to answer the evaluated questions and answers, per language

In general, with the exception of Bulgarian and Spanish language, most evaluators present low or medium knowledge of the specific subject of the question. This level of knowledge is typical of the real customers of the PcMedic Wizard application and makes the results of the evaluation relevant to reflect the real usage scenario.

5.3 Usefulness

Tables 29, 30 and 31 shows the evaluation results based on Form 2, where the evaluator is asked to assess on the usefulness of the automatically translated answers by Pilot 0 and Pilot 3.

Regarding the Basque language, 30% of the the answers delivered by Pilot 3 were considered not helpful or understandable, in comparison with the 13% of the answers translated with Pilot 0 in such conditions.

For the Bulgarian language, Pilot 3 outperforms Pilot 0, with 71% of answers translated with Pilot-3 considered clearly helpful, against 28% by Pilot 0; and with only 5% of answers translated by Pilot 3 considered not helpful/understandable, against 9% by Pilot 0.

German Pilot 3 and Pilot 0 obtained very similar results: 26% of the answers for both Pilots were considered not helpful/understandable. However Pilot 3 obtained a higher percentage of clearly helpful answers than Pilot 0 (33% against 30%).

When we look at the results for the Czech language, Pilot 3-Chimera outperforms Pilot 0, as the percentage of not helpful/understandable answers are lower (44% against 48%), while Pilot 3-TectoMT presents results slightly worse than Pilot 0.

It is worth noting that, the cohort of evaluators assessing Pilot 3-Chimera (vs. Pilot 0) might be different from the cohort of evaluators assessing Pilot 3-TectoMT (vs. Pilot 0). Additionally, the set of sentences was the same for both experiments but the order in which the sentences were presented could be different. So if the same annotator evaluates a subset of 25 sentences from the set relevant for Pilot 3-TectoMT and 25 sentences from the set relevant for Pilot 3-Chimera, it is likely that he did not annotate the same sentences. It is also worth noting that even if the cohort happened to be the same and in the rare cases where the same annotators annotate the same sentences in the two

	EU		BG		DE	
	P0	P3	P0	P3	P0	P3
It would clearly help me solve my problem / answer my question	45%	22%	28%	71%	30%	33%
It might help, but would require some thinking to understand it.	41%	48%	64%	24%	44%	41%
Is not helpful / I don't understand it	13%	30%	9%	5%	26%	26%

Table 29: Assessment of the usefulness of the translated answers for Basque, Bulgarian and German

	CS				NL			
	P0	P3	P0	CH	P0	P3	P0	CH
It would clearly help me solve my problem / answer my question	26%	22%	20%	19%	21%	24%	27%	31%
It might help, but would require some thinking to understand it.	35%	38%	32%	38%	49%	47%	43%	42%
Is not helpful / I don't understand it	39%	40%	48%	44%	30%	30%	30%	27%

Table 30: Assessment of the usefulness of the translated answers for Czech and Dutch

	PT				ES			
	P0	P3	P0	CH	P0	P3	P0	CH
It would clearly help me solve my problem / answer my question	3%	6%	8%	26%	56%	54%	42%	63%
It might help, but would require some thinking to understand it.	30%	56%	31%	56%	35%	39%	41%	28%
Is not helpful / I don't understand it	67%	38%	61%	18%	9%	8%	17%	9%

Table 31: Assessment of the usefulness of the translated answers for Portuguese and Spanish

experiments, it is only natural and expected that the laypersons in that cohort may rate some questions differently in different sessions (without having access to their previously assigned scores). This is the reason why there are two columns with scores for the Pilots 0 (for Czech, Dutch, Portuguese and Spanish) that are not necessarily identical.

Dutch Pilot 3-TectoMT and Pilot 0 obtained very similar results as 30% of the answers for both Pilots were considered not helpful/understandable. However Pilot 3-TectoMT obtained a higher number of clearly helpful answers than Pilot 0, 24% against 21%. Pilot 3-Chimera obtained better results in terms of both the percentage of not useful/understandable answers (27% versus 30%) and the percentage of clearly helpful (31% versus 27%).

Regarding the Portuguese language, Pilot 3-TectoMT outperforms Pilot 0 in all the dimensions and Pilot 3-Chimera increased further the improvement over Pilot 0.

Spanish Pilot 3-TectoMT is slightly better than Pilot 0 in terms of percentage of answers considered as not helpful/understandable, while Pilot 3-Chimera obtained better scores, increasing the percentage of answers considered clearly helpful (from 42% to 63%) and lowering the percentage of answers considered not helpful/understandable (from 17% to 9%).

In general, Pilot 3-TectoMT or Pilot 3-Chimera improved over Pilot 0 for almost all languages, except Basque. In some cases, the improvement consisted of a few points, as

	Average			
	P0	P3	P0	Chimera
It would clearly help me solve my problem / answer my question	30.0%	32.8%	29.5%	37.7%
It might help, but would require some thinking to understand it.	43.4%	42.0%	42.9%	39.9%
Is not helpful / I don't understand it	26.6%	25.2%	27.6%	22.4%

Table 32: Average of the assessment of the usefulness of the translated answers across languages

in the case of German, or was quite large, as in the case of Bulgarian and Portuguese.

Table 32 reports the average assessment of the usefulness of the answer for all the languages. On average, a system where Pilot 3 is used gets a more positive evaluation than a system where Pilot 0 is used. A system where Chimera is used (for the languages where this system is available) obtains yet more positive judgments.

5.4 Correctness

Tables 33, 34 and 35 report on the results from Form 3, where the evaluator was asked to compare the answers automatically translated by Pilot-0 and by Pliot-3 with the reference answer, taking into consideration that the reference answer gives the correct information.

	EU		**BG**		**DE**	
	P0	P3	P0	P3	P0	P3
gives the right advice	39%	12%	24%	67%	41%	38%
gets minor points wrong	42%	43%	64%	25%	35%	35%
gets major points wrong	19%	45%	11%	7%	23%	27%

Table 33: Assessment of the correctness of the translated answers, when judged against the reference answers, for Basque, Bulgarian and German

	CS				**NL**			
	P0	P3	P0	CH	P0	P3	P0	CH
gives the right advice	19%	18%	14%	16%	25%	27%	25%	26%
gets minor points wrong	38%	39%	35%	33%	37%	42%	41%	41%
gets major points wrong	43%	43%	52%	52%	38%	30%	34%	33%

Table 34: Assessment of the correctness of the translated answers, when judged against the reference answers, for Czech and Dutch

	PT				**ES**			
	P0	P3	P0	CH	P0	P3	P0	CH
gives the right advice	2%	6%	8%	24%	55%	51%	29%	50%
gets minor points wrong	30%	54%	24%	51%	26%	27%	43%	33%
gets major points wrong	68%	40%	67%	25%	19%	22%	28%	18%

Table 35: Assessment of the correctness of the translated answers, when judged against the reference answers, for Portuguese and Spanish

These results are obtained when the evaluators are presented with the reference answer. In general they are in line with the results obtained with Form 2 (displayed in previous Tables 29, 30 and 31) with a slight worsening. Here, evaluators rating of the automatic

	Average			
	P0	P3	P0	Chimera
gives the right advice	29.3%	30.8%	26.4%	32.9%
gets minor points wrong	39.8%	38.3%	41.4%	37.4%
gets major points wrong	30.8%	31.0%	32.2%	29.7%

Table 36: Average of the assessment of the correctness of the translated answer, judged against the reference answer, across languages

translated answer gets a little worse. The only partial exception is German, where the percentage of answers considered as giving the right advice is higher than the percentage of the answers considered clearly helpful in Form 2 for both Pilots.

For the Bulgarian, Dutch and Portuguese languages, the results from Form 3 does not change the overall evaluation, where Pilot 3-TectoMT or Pilot 3-Chimera outperforms Pilot 0.

Regarding Czech, the evaluation obtained by Pilot 3-Chimera with Form 2 is cut down as both Pilot 0 and Pilot 3-Chimera obtain here the same percentage of answers considered as getting major points wrong (52%). However Chimera gets a higher percentage of answers evaluated as getting the right advice (16% against 14%).

Finally Spanish Pilot 0 outperforms Pilot 3-TectoMT, but Pilot 3-Chimera still outperforms Pilot 0.

Table 36 presents the overall results when considering all the languages. Regarding this evaluation, there is no clear supremacy of Pilot 3 over Pilot 0, while Chimera remains the best option.

5.5 Probability of calling an operator

In deliverable D3.3, a metric was proposed that aims at determining the probability of a final user to make a phone call in order to get a satisfactory answer to his questions. What is relevant for this metric is the perception of the user about the usefulness of the answer. This means that if in Form 2 the evaluator checked that the automatically translated answer "clearly help(s) to solve my problem / answer(s) my question", the probability that he will ask for further (telephone) help is very low. This is especially expected if the answer, when compared to a reference answer, is judged as giving "the right advice" or having just "minor points wrong".

The case is different when an evaluator thinks that the (translated) answer "would require some thinking to understand it" and "gets important points wrong". In this case, for example, the probability of calling an operator is high. Taking in consideration the answers given in Form 2 and Form 3, a ranking metric was elaborated. Table 37 displays the three-level probability of calling an operator for each different possibility in the checking of Forms 2 and 3 along this metric.

The results for each language are presented in Tables 38, 39 and 40 where the scores indicate the proportion of answers per case, from *A* to *I*.

In order better help to draw some conclusions, the aggregates results are presented in Tables 41, 42 and 43

Regarding Basque, the probability of calling an operator to get further help is higher for Pilot 3 than for Pilot 0.

For Bulgarian, in 74% of the cases there is a small probability to call an operator when the answers where translated with Pilot 3. This represents an improvement over Pilot 0,

	Form2	Form3	Probability
A	Solves my problem	Gets the right advice	**low**
B	Solves my problem	Gets minor points wrong	**low**
C	Would require some thinking to understand it	Gets the right advice	**low**
D	Would require some thinking to understand it	Gets minor points wrong	**medium**
E	Solves my problem	Gets important points wrong	**high**
F	Would require some thinking to understand it	Gets important points wrong	**high**
G	Is not helpful / I don't understand it	Gets the right advice	**high**
H	Is not helpful / I don't understand it	Gets minor points wrong	**high**
I	Is not helpful / I don't understand it	Gets important points wrong	**high**

Table 37: The metric for estimating the probability of calling an operator

		EU		BG		DE	
		P0	P3	P0	P3	P0	P3
A	**low**	21%	11%	19%	38%	22%	24%
B	**low**	18%	10%	11%	29%	6%	7%
C	**low**	4%	11%	25%	7%	14%	12%
D	**medium**	22%	25%	34%	17%	22%	21%
E	**high**	7%	2%	0%	2%	2%	2%
F	**high**	15%	12%	2%	2%	8%	8%
G	**high**	0%	3%	1%	0%	4%	3%
H	**high**	3%	8%	2%	0%	7%	7%
I	**high**	10%	19%	5%	3%	14%	15%

Table 38: Estimated probability of calling and operator for Basque, Bulgarian and German: breakdown per nine levels of probability

where the probability of calling an operator is low only in 55% of the cases.

The probabilities are very similar between Pilot 0 and Pilot 3 for German, with Pilot 3 presenting a slight improvement in terms of the questions with low probability of triggering a call to the operator for further help.

For Czech, Pilot 0, Pilot 3-TectoMT and Pilot 3-Chimera do not differ much along the dimension being assessed here.

For Dutch, there is an improvement regarding the answers with low probability of leading to call to an operator for both Pilot 3-TectoMT and Pilot 3-Chimera, while the proportion of questions in the high probability class is the same.

For the Portuguese language, both Pilot 3-TectoMT and Pilot 3-Chimera improve on Pilot 0. The best results are obtained by Pilot 3-Chimera, that increases the proportion of questions with low probability of calling an operator up to 22%, from the 13% found for Pilot 0, and decreases the high probability class of questions down to 49%, from the score of 69% of Pilot 0.

For Spanish, Pilot 0 and Pilot 3-TectoMT present quite similar scores, while Pilot 3-Chimera outperforms Pilot 0 with an improvement in both classes of low and high probability of calling an operator.

The scores of the *A* to *I* cases in Tables 38, 39 and 40, and their aggregated rendering in Tables 41, 42 and 43 are obtained from the scores in Forms 2 and 3, and are thus in line with the later. In general, Pilot 3-TectoMT or Pilot 3-Chimera improved over Pilot 0 for almost all languages, except Basque. In some cases, the improvement consisted of a few points, as in the case of German, or are substantial, as in the case of Bulgarian or Portuguese.

		CS				NL			
		P0	P3	P0	CH	P0	P3	P0	CH
A	**low**	12%	12%	10%	11%	10%	11%	15%	17%
B	**low**	12%	9%	8%	6%	8%	10%	10%	12%
C	**low**	3%	5%	5%	4%	12%	12%	8%	6%
D	**medium**	21%	21%	19%	21%	23%	21%	27%	26%
E	**high**	1%	1%	2%	2%	4%	3%	0%	1%
F	**high**	12%	12%	9%	14%	14%	15%	9%	10%
G	**high**	3%	2%	0%	1%	5%	4%	1%	1%
H	**high**	7%	9%	7%	7%	9%	10%	4%	4%
I	**high**	30%	30%	41%	36%	16%	16%	25%	23%

Table 39: Estimated probability of calling and operator for Czech and Dutch: breakdown per nine levels of probability

		PT				ES			
		P0	P3	P0	CH	P0	P3	P0	CH
A	**low**	1%	2%	6%	11%	38%	37%	27%	35%
B	**low**	1%	2%	2%	7%	11%	10%	14%	22%
C	**low**	0%	2%	6%	4%	13%	15%	10%	4%
D	**medium**	22%	33%	18%	29%	13%	14%	21%	14%
E	**high**	0%	2%	1%	8%	6%	7%	1%	6%
F	**high**	7%	22%	8%	24%	10%	10%	10%	10%
G	**high**	2%	0%	5%	1%	2%	1%	2%	0%
H	**high**	19%	7%	18%	2%	2%	2%	3%	1%
I	**high**	46%	31%	38%	15%	6%	5%	12%	8%

Table 40: Estimated probability of calling and operator for Portuguese and Spanish: breakdown per nine levels of probability

	EU		BG		DE	
	P0	P3	P0	P3	P0	P3
low	43%	32%	55%	74%	42%	43%
medium	22%	25%	34%	17%	22%	21%
high	35%	43%	11%	8%	36%	36%

Table 41: Estimated probability of calling and operator for Basque, Bulgarian and German: results aggregated by three levels of probability

	CS				NL			
	P0	P3	P0	CH	P0	P3	P0	CH
low	27%	25%	23%	20%	29%	32%	33%	36%
medium	21%	21%	19%	21%	23%	21%	27%	26%
high	53%	53%	59%	59%	48%	48%	39%	39%

Table 42: Estimated probability of calling and operator for BCzech and Dutch: results aggregated by three levels of probability

	PT				ES			
	P0	P3	P0	CH	P0	P3	P0	CH
low	3%	6%	13%	22%	62%	61%	51%	61%
medium	22%	33%	18%	29%	13%	14%	21%	14%
high	75%	61%	69%	49%	25%	24%	28%	24%

Table 43: Estimated probability of calling and operator for Portuguese and Spanish: results aggregated by three levels of probability

	Average			
	P0	P3	P0	CH
low	37.3%	39.1%	37.2%	41.2%
medium	22.4%	21.7%	23.2%	21.9%
high	40.3%	39.2%	39.6%	36.9%

Table 44: Average of the estimated probability of calling an operator across languages: results aggregated by three levels of probability

	EU	**BG**	**CS**	**CS^C**	**NL**	**NL^C**	**DE**	**PT**	**PT^C**	**ES**	**ES^C**
a. P0 better than P3	60%	12%	27%	26%	31%	15%	31%	16%	11%	41%	11%
b. P3 better than P0	11%	58%	25%	28%	39%	26%	28%	45%	61%	26%	40%
c. Equally good	16%	22%	16%	13%	12%	27%	25%	6%	9%	12%	31%
d. Equally bad	12%	7%	31%	34%	19%	32%	17%	32%	19%	21%	17%
e. P0 as good as P3 (c+d)	28%	30%	48%	47%	31%	59%	42%	38%	28%	33%	49%
f. P3 better ignoring ties	16%	83%	48%	52%	56%	64%	47%	74%	85%	39%	78%
g. significant-McNemar	yes	yes	no	no	no	yes	no	yes	yes	yes	yes

Table 45: Comparison between Pilot 0 and Pilot 3 translations. CS, NL, PT and ES, and CS^C, NL^C, PT^C and ES^C, are respectively the columns for Pilot3-TectoMT and for Pilot3-Chimera results.

As shown in Table 44, containing the average of the estimated probabilities across languages, there is an improvement in the publication step from Pilot-0 to Pilot-3: there is a gain of almost 2 percentage points with Pilot-3-TectoMT and of 6 percentage points with Pilot-3-Chimera.

5.6 Human intrinsic evaluation

Table 45 presents the results collected with Form 4, where the evaluators were asked to perform intrinsic evaluation by comparing Pilot 0 and Pilot 3 translations EN→X on their own merit as translations, that is with respect to the gold-standard translation and disregarding any concern with their impact to the QA usage scenario where they may be embedded.

As we can see in row *e*, the percentage of ties differs across languages. Therefore, we cannot compare the relative quality of Pilot 3 and Pilot 0 only based on the number of cases when Pilot 3 was judged strictly better than Pilot 0 (row *a*). Thus, in the row *f*, we report the percentage of non-tying comparisons where Pilot 3 was judged better than Pilot 0., i.e.:

$$row\ f = P3\text{-}better\text{-}ignoring\text{-}ties\text{-}than\text{-}P0$$

$$row\ f = \frac{\#(P3\text{-}better\text{-}than\text{-}P0)}{\#(P3\text{-}better\text{-}than\text{-}P0) + \#(P3\text{-}worse\text{-}than\text{-}P0)} \times 100\%$$

$$row\ f = \frac{row\ b}{row\ b + row\ a} \times 100\%$$

Row *g* shows whether the difference between Pilot 0 and Pilot 3 (or Pilot3-Chimera) is significant according to McNemar's test at the 95% confidence level.

Figure 13 provides a graphical representation of Table 45, where the Chimera version of Pilot 3 was selected for the languages where it is available. The languages (vertical

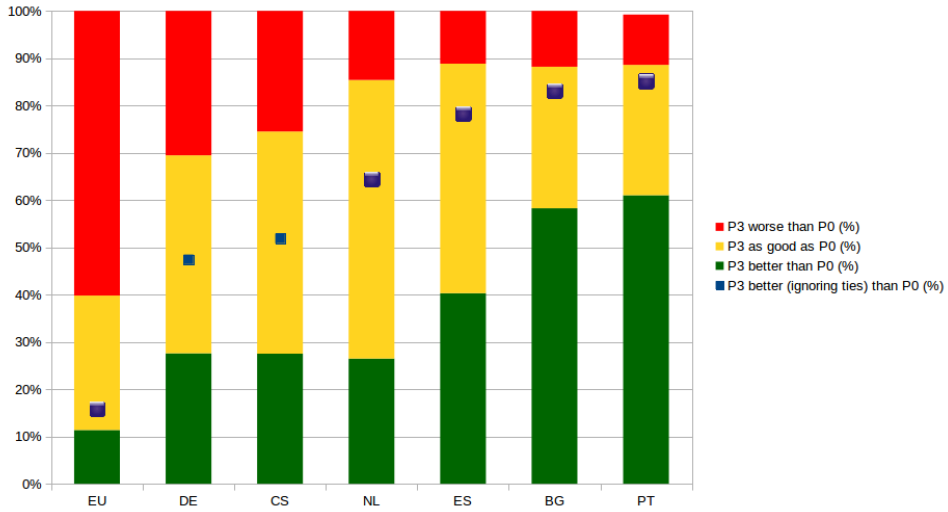


Figure 13: Breakdown of the human evaluation for the comparison between Pilot 3 and Pilot 0, per language, where the Chimera version of Pilot 3 was selected for the languages where it is available, viz. CS, ES, NL and PT

bars) in the figures are sorted by the *better ignoring ties* scores, which are plotted as dark blue boxes. Bigger boxes indicate significant differences (row g). Note that we treat *P3 better than P0* as equivalent to *P0 worse than P3*.

In general, the intrinsically and manually assessed quality of translations by Pilot 3 are as good as or better than the translations by Pilot 0, the baseline, for all languages except Basque.

Basque is a less-resourced language and though the QTLeap project has contributed greatly to improve on this situation with the datasets and processing tools curated for this language, the results obtained for the MT Pilots addressing Basque should be understood as a further stimulus to keep looking to improve the language technology for this language.[4]

Given this circumstance for Basque, for a vast majority of two thirds of the language pairs in the project, the baseline Pilot 0 has been significantly, and in some cases substantially, outperformed by the machine translation technology resorted to and enhanced in this project in terms of manual intrinsic evaluation: for these languages, the *better ignoring ties* score — represented by level of the blue square — is significantly higher than 50%.

5.7 Inter-annotator agreement

The inter-annotator agreement is presented in Tables 46, 47, 48 and 49 as the ratio of the times two annotators agree with each other to the total amount of pairwise comparisons. The inter-annotator agreement is not comparable to Pilots 1 and 2, as this type of evaluation did not take place for them.

The inter-annotator agreement is relatively low, as annotators agree with each other

[4]A Pilot 3-Chimera was developed also for Basque and it scores significantly better than Pilot 3 in terms of BLEU (17.16 vs. 11.24, see D2.11). Unfortunately, there was not enough human resources to evaluate it manually in the evaluation reported in the present deliverable. Thus Basque Pilot 3-Chimera is not evaluated in this deliverable and only Basque Pilot 3-TectoMT is reported on.

lang.	annotators	agreement
bg	5	0.504
cs	3	0.430
de	7	0.358
es	3	0.497
eu	3	0.573
nl	11	0.364
pt	3	0.544

Table 46: Inter-annotator agreement per language pair for Pilot 0 against Pilot 3

task	agreement
Form2 Pilot 0	0.537
Form2 Pilot 3	0.503
Form3 Pilot 0	0.511
Form3 Pilot 3	0.494
Form4	0.374

Table 47: Inter-annotator agreement per task for Pilot 0 against Pilot 3

lang.	annotators	agreement
cs	2	0.45
es	5	0.432
nl	19	0.341
pt	4	0.561

Table 48: Inter-annotator agreement per language pair for Pilot 0 against Chimera

task	agreement
Form2P0	0.518
Form2P3	0.518
Form3P0	0.531
Form3P3	0.482
Form4	0.343

Table 49: Inter-annotator agreement per task for Pilot 0 against Chimera

almost half of the times.

In terms of comparing between language pairs, the best inter-annotator agreement was achieved for Basque and Portuguese, whereas the lowest was achieved for German and Dutch, a fact that can be partially justified by the absolute value of difference in quality between the pilots (cf. Figure 14).

In terms of comparing between the tasks, forms 2 and 3 demonstrate similar inter-annotator agreement, whereas Form4 indicates less agreement.

There is no considerable difference in inter-annotator agreement for these comparisons when it comes to comparing Pilot 0 with Chimera.

	EU	BG	CS	CS^C	NL	NL^C	DE	PT	PT^C	ES	ES^C
P0 BLEU	17.94	20.30	23.17	23.17	25.42	25.42	34.90	12.01	12.01	24.11	24.11
P3 BLEU	11.24	23.91	24.24	26.16	22.35	26.65	31.12	15.33	19.64	24.94	35.36
ΔBLEU	−6.70	3.61	1.07	2.99	−3.07	1.23	−3.78	3.32	7.63	0.83	11.25

Table 50: Difference between the Pilot 0 and Pilot 3 in terms of BLEU scores, per language.

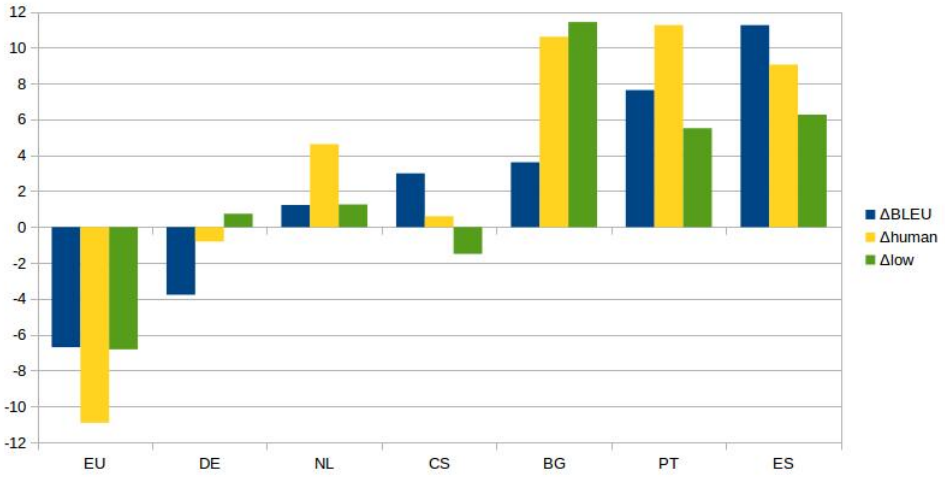


Figure 14: Comparison of the extrinsic evaluation results with intrinsic evaluation results in the publication step, per language: Comparison of the differences between Pilot 0 and Pilot 3 in terms of their BLEU scores, the human evaluation and the low probability to call an operator, per language, where the Chimera version of Pilot 3 was selected for the languages where it is available, viz. CS, ES, NL and PT. Y-axis is in BLEU points and the top bars for the other two metrics (representing percentage points) were re-scaled to align with the top bar for the BLEU metric.

5.8 Comparison with intrinsic evaluation

This human intrinsic evaluation data was compared with the automatic intrinsic performance measured by BLEU.

Table 50 shows the BLEU scores of the two Pilots and the differences between them, as described in detail in D2.11.

Figure 14, in turn, presents the BLEU difference (ΔBLEU, dark blue bars) in relation to the difference between Pilots according to the human evaluation (Δhuman, yellow bars), again selecting Chimera where available. It also presents the comparison with the probability to call an operator, where this probability is low (Δlow, green bars).

For the purpose of Figure 14, we have computed Δhuman as the *better ignoring ties* score scaled to the same range as ΔBLEU (keeping 0 as the neutral point and setting $\max \Delta human = \max \Delta BLEU$), which boils down to

$$\Delta human = (P3\text{-}better\text{-}ignoring\text{-}ties\text{-}than\text{-}P0 - 50\%) \cdot constant$$
$$constant = (\max P3\text{-}better\text{-}ignoring\text{-}ties\text{-}than\text{-}P0 - 50\%)^{-1} \cdot (\max \Delta BLEU)$$

The languages are represented by clusters of bars in Figure 14, which are sorted according to increasing values for ΔBLEU.

As expected from previous studies on the correlation between BLEU and human evaluation of translation quality, the width of the gaps between Pilot 3 and Pilot 0 as these are measured automatically with BLEU are not proportional to the width of such gaps as they are measured manually. Against this background, what is worth noting is that the signal of the difference between the performance scores of Pilot 3 vs. its baseline Pilot 0 as these are assessed automatically for intrinsic evaluation shows and overall alignment with the signal of the respective difference as this is assessed manually, thus providing a further piece of reciprocal confirmation for their reliability.

As for the comparison between extrinsic and intrinsic evaluation results, in general, the signals of the Δs are similar for all languages except for Czech (with positive intrinsic Δ yet with negative extrinsic Δ) and German (with negative intrinsic Δ yet with slightly positive extrinsic Δ)), thus indicating that an improvement in terms of MT performance, from the baseline Pilot-0 to the Pilot 3, tends to induce a consistent improvement in terms of the publication step of the QA system.

6 Calculation of time saving/cost reduction

In order to asses the impact of the different Pilots on its business, HF has performed a rough calculation of the costs for migrating their service to a different language based on the given MT technology. From the HF business perspective, it is hard at this point to draw an estimate of cost benefit when MT is used, as at the moment the company works only with Portuguese/English without any type of translation (human or automatic). As we have seen in both the retrieval and publication steps, Pilot 3 leads to better results than Pilot 0 for most of the languages, both when intrinsically or extrinsically evaluated. We want to understand in which measure this improvement can impact the economy of the QA PcWizard application.

We contextualized the impact of the automatic translation service on the PcWizard application taking in account the undergoing process of internationalization carried out by HF. In this context, we need to consider the cost of developing and driving the product effectively into a new country. While the ultimate goal is to arrive at an MT-supported fully automatic scenario for any new language, we are presenting a calculation for a more conservative post-editing scenario below where the HF database needs to be populated for a new language.

We assume that foreign-language operators are available. As the business model is to answer recurrent questions, the effort needed for post-editing an automatically translated question once (the correction will then go into the database) can be neglected here. Actually, the same calculation can be applied in the presence of a multi-lingual system, where new questions come from clients and new answers are created. In this case, when a new interaction in a particular language is introduced in the system, the QA algorithm can make use of it to answer a question in another language.

6.1 Extrinsic evaluation: combining the retrieval and the publication steps

We tried to estimate the combined impact of the retrieval and publication step as the probability of answering a question without resorting to any human intervention.

We made this estimation for both Pilot 0 and Pilot 3.

	EU	BG	CS	NL	DE	PT	ES
AccuracyPSL >=75 P0	19.6%	21.0%	25.3%	29.1%	33.0%	17.4%	25.9%
AccuracyPSL >=75 P3	19.0%	27.4%	34.6%	32.8%	11.5%	22.3%	28.0%
Probability low P0	43.1%	55.1%	22.8%	33.4%	41.9%	13.1%	51.0%
Probability low P3	31.8%	74.2%	20.3%	35.5%	43.1%	22.2%	61.4%
Fully automatic P0	8.5%	11.6%	5.8%	9.7%	13.8%	2.3%	13.2%
Fully automatic P3	6.0%	20.3%	7.0%	11.6%	5.0%	5.0%	17.2%

Table 51: Overall extrinsic evaluation, combining retrieval and publication steps (table): Proportion of questions that can be answered without human intervention, i.e. fully automatically, by Pilot 0 and by Pilot 3, per language, where the Chimera version of Pilot 3 was selected for the languages where it is available, viz. CS, ES, NL and PT, and where figures for DE reflect the impact of having experimented with neural MT technology for P3 with this language..

For the sake of this estimation, we considered the lower threshold ($\geq$ 75) of answers that are showed directly to the user without human intervention. This decision is based on the observation of the accuracy per ranking position of the first candidate answer (AccuracyPRP@1), where it is evident that, on overage, the average score of correct answers, when MT is used, is lower than the one when the original English is used. For this reason we assumed that the answers with a score $\geq$75 would be presented directly to the user.

Among such answers, only those that are correct will not generate a further request for help by the customer. This corresponds to the percentage of accurate answers with scoring level $\geq$75 (AccuracyPSL$\geq$75).

We know that even correct answers have a residual probability to generate a call for help as they may not be clear. If we multiply the percentage of answers with the accuracy score of $\geq$75 with the percentage of answers that present a low probability of calling an operator, we can estimate the percentage of answers that can be attended without human intervention, that means in a fully automatic way.

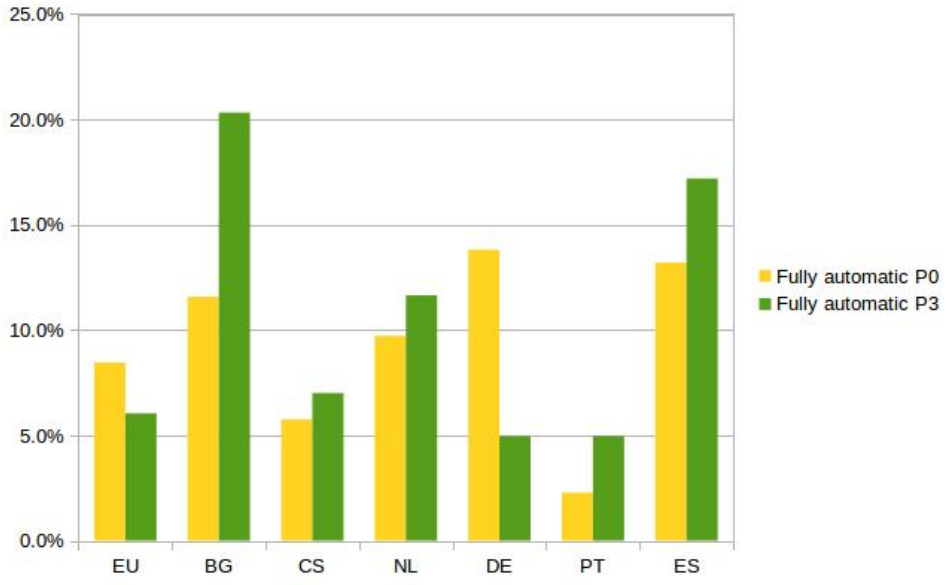


Figure 15: Overall extrinsic evaluation, combining retrieval and publication steps (synopsis chart): Proportion of questions that can be answered without human intervention by Pilot 0 and by Pilot 3, per language, where the Chimera version of Pilot 3 was selected for the languages where it is available, viz. CS, ES, NL and PT, and where figures for DE reflect the impact of having experimented with neural MT technology for P3 with this language.

	EU	BG	CS	NL	DE	PT	ES
Δ Fully automatic	−2.5	8.7	1.2	1.9	−8.8	2.7	4.0

Table 52: Overall extrinsic evaluation, combining retrieval and publication steps (table with Δs): summary of progress from Pilot 0 to Pilot 3 (details in Table 51), per language, in percentage points, where the Chimera version of Pilot 3 was selected for the languages where it is available, viz. CS, ES, NL and PT, and where figures for DE reflect the impact of having experimented with neural MT technology for P3 with this language.

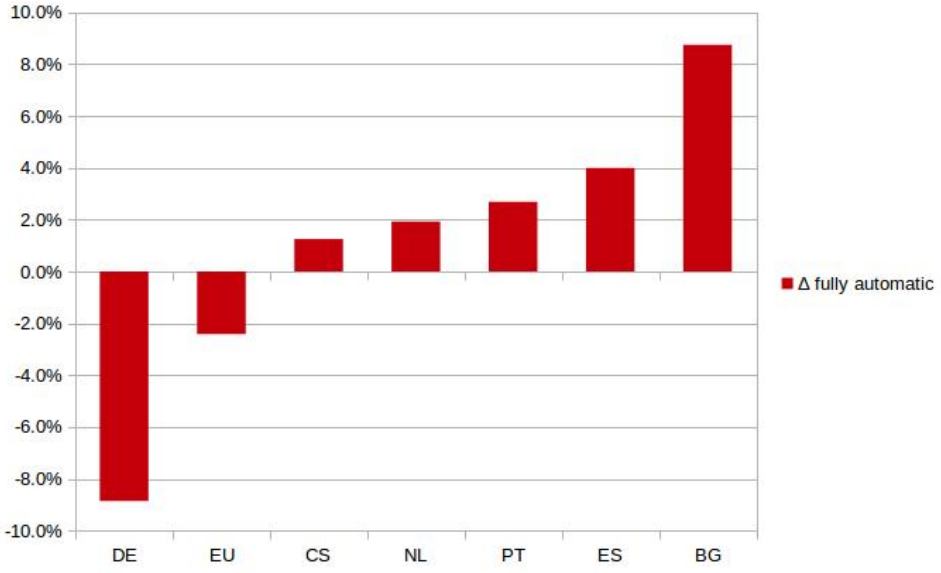


Figure 16: Overall extrinsic evaluation, combining retrieval and publication steps (chart with Δs): summary of progress from Pilot 0 to Pilot 3 (details in Table 51), per language, in percentage points, where the Chimera version of Pilot 3 was selected for the languages where it is available, viz. CS, ES, NL and PT, and where figures for DE reflect the impact of having experimented with neural MT technology for P3 with this language.

Table 51 and companion Figure 15 show the results of this overall extrinsic evaluation. For all the languages there is a consistent gain in using a MT system to deal with a new language.

For some languages the percentage of answers that can be returned without human intervention is quite impressive. For Bulgarian, for example this percentage can exceed 20% and for Spanish 17%. Figure 15 gives a graphical view of the performance of the QA systems with the embedding of the two MT Pilots.

Regarding the progress from Pilot 0 to Pilot 3, the results can be more easily appreciated in Table 52 and its companion Figure 16, with figures representing the difference between the extrinsic evaluation results of the QA systems.

The QA systems with Pilot 3 embedded in them present better results for most of the languages. The difference can be substantial as in the case of Bulgarian with a improvement of almost 9 percentage points.

An exception is German, whose figures reflect the impact of having experimented with neural MT technology for P3. And the other exception is the Basque language, due to the reasons discussed above related to it being under resourced in terms of language resources and tools.

6.2 Comparison with intrinsic evaluation

Additionally, it is important to understand whether higher quality MT systems can substantially impact the performance of the QA system and to what extent. It is thus important finally to compare the progress of the MT systems, in terms of intrinsic evaluation results, with the progress of the QA systems with these MT systems embedded in them, in terms of the overall extrinsic evaluation results.

A synopsis of this comparison exercise is presented in the chart of Figure 17, which brings together, per language, the differences from Pilot 0 to Pilot 3 in terms of the overall extrinsic evaluation score and in terms of the intrinsic evaluation scores.

The intrinsic scores concern the retrieval step and the publication step.

The score for the retrieval step are in BLEU points and come from Table 24 (and companion Figure 5).

The scores for the publication step are in BLEU points for the automatic evaluation and come from Table 50, and are in percentage points for the human evaluation and come from Table 45 (and companion Figure 13). Figure 14 compares the scores from these two intrinsic evaluation procedure for the publication step.

The extrinsic scores concerns the overall process of getting an automatically returned answer upon the corresponding question have been input. They are in percentage points and come from Table 52 (and companion Figure 16).

The top bars for the metrics in percentage points were re-scaled to align with the top bar for the BLEU metric (scoring 11.25 BLEU points for Spanish, cf. Table 50 and companion Figure 14), along the lines adopted above for the intrinsic evaluation metrics of the publication step presented in Figure 14.

The signal of the extrinsic evaluation outcome is identical to the signal of the outcome of all the three intrinsic evaluation metrics for most of the languages, except for Portuguese and Spanish.

For Portuguese and Spanish, the signal (positive) of the extrinsic evaluation is identical to the signal of two of the intrinsic evaluation metrics, namely the two metrics concerning the publication step, and is opposite to the signal of the intrinsic evaluation metric for the retrieval step (which is negative). The circumstance that there was progress from Pilot 0 to Pilot 3 in the outbound direction (EN→X) and there was no such progress for the inbound direction (X→EN) was not enough to overcome the larger, beneficial effect of the first and to cancel a substantial positive impact in the progress of the QA system from Pilot 0 to Pilot 3, as indicated by the positive outcome of the overall extrinsic evaluation.

There is substantial progress in the performance of the QA systems in terms of providing fully automatic answers from their versions that have Pilots 1 embedded in them to the version that include Pilots 3 for most languages except Basque and German.

The reasons for the results concerning Basque and German have been thoroughly explained above and relate to Basque being a less resourced language in terms of previously available data resources and processing tools and to German has been opted for concerning the comparative experimentation with the alternative neural machine translation technology.

The progress of the overall QA systems that have the MT Pilots embedded in them can reach up almost 9 percentage points in the proportion of answers that can be returned fully automatically without the intervention of a human operator (in the case of Bulgarian), thus almost doubling the proportion of questions that can be answered fully automatically, by bringing this proportion to 20% of all entered questions. This is the consequence of the progress achieved in terms of machine translation, where the deep MT Pilot 3 can reach

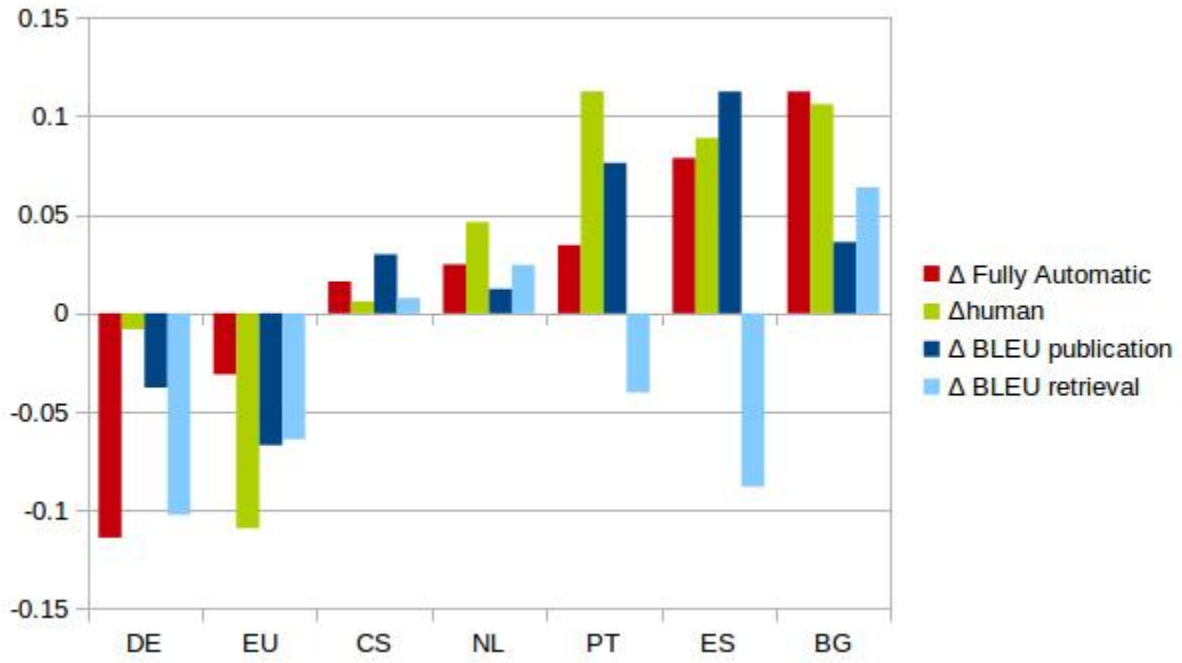


Figure 17: Comparison of intrinsic evaluation outcome in both retrieval and publication steps with overall extrinsic evaluation outcome: Comparison of the differences between Pilot 0 and Pilot 3 in terms of the results from intrinsic evaluations of the translation directions X→EN (Δ BLEU retrieval) and EN→X (Δ BLEU publication and Δ Human), and of the results from overall extrinsic evaluation (Δ Fully Automatic), per language. Y-axis is in BLEU points and the top bars for the other two metrics (representing percentage points) were re-scaled to align with the top bar for the BLEU metric (scoring 11.25 BLEU points for Spanish, cf. Table 50 and companion Figure 14).

up to 85% probability of delivering a better translation (ignoring ties) than the baseline SMT Pilot 0 in the EN→X translation direction according to human evaluators (in the case of Portuguese).

7 Conclusion

This evaluation focused on the impact of the translation delivered by Pilot 3 on the QA PcMedic Wizard, a helpdesk application developed by the industrial partner HF as part of its business. In general, the focus of this evaluation was to assess the added value of the translations in terms of their impact on the performance of the QA system of the helpdesk. In this particular evaluation, we compared the results of Pilot 3 with the results of the baseline Pilot 0. We extended this comparison to all the other Pilots, when automatic evaluation instead of human evaluation was carried on.

The extrinsic evaluation was divided in two parts. The first one aimed to test the impact of the translation system on the retrieval step of the QA PcMedic Wizard. In this case, the "source language" → English translation direction was taken into account. The main goal of this part was to compare the result obtained when the original English question is used by the QA algorithm with the result obtained when the question was translated from a different language to English. As this evaluation is automatic and does not require human intervention all the Pilots were tested.

We measured the answering rate, the percentage of how many questions the QA algorithm was able to find a candidate answer to within a certain confidence score interval. Better results were obtained with Pilot 3 translations than with Pilot 0 for all languages but German. For one language, Czech, Pilot 2 slightly outperformed Pilot 3, while for Portuguese and Spanish these two Pilots obtained very similar performances.

Regarding the accuracy per ranking positions in the retrieval step, Pilot 3 scored better in comparison with Pilot 0 and also the other Pilots for most languages, namely Bulgarian, Czech, Dutch and Portuguese (with the exception of Czech where Pilot 2 and Pilot 3 were very similar).

The second part of the extrinsic evaluation targeted the publication step of the QA system and was carried out by human evaluators. In this step, the English → "target language" translation direction was taken into account. Pilot 0 and Pilot 3 have been tested by volunteer subjects. For four languages, namely Czech, Dutch, Portuguese and Spanish a fifth translation system was evaluated named Pilot 3-Chimera, resulting from the combination of TectoMT and Moses systems.

The metric presented in deliverable D3.3 was used to asses the probability to call an operator, with lower the probability better the contribution of the MT system. If we consider only the level where the probability of calling an operator is low, better results were obtained when Pilot 3-TectoMT or Pilot 3-Chimera were used for all languages but Basque and Czech.

The present evaluation exercise included also a part on manual intrinsic evaluation. Pilot 0 and Pilot 3 translations of answers were also directly compared by human evaluators. The result of this comparison is that the Pilot 3-TectoMT or Pilot 3-Chimera answers are significantly better than Pilot 0 for most languages in the project, namely Bulgarian, Dutch, Portuguese and Spanish. For German and Czech there is no significant difference between the two Pilots and only Basque Pilot 3 is worse than Pilot 0.

The impact of the MT systems with increasingly higher quality on the overall QA systems, including both retrieval and publication steps, was also assessed.

As thoroughly documented in this deliverable, the progress of the QA systems that have the MT Pilots embedded in them can reach up almost 9 percentage points in the proportion of answers that can be returned fully automatically without the intervention of a human operator, thus almost doubling the proportion of such questions, by bringing this proportion to 20% of all entered questions. This is the consequence of the progress achieved in terms of machine translation, where the deep MT Pilot 3 can reach up to 85% probability of delivering a better translation than the baseline SMT Pilot 0 direction according to human evaluators.

Even though these results obtained on a sample of 100 interactions should be interpreted with care, as has been underlined, it is worth noting that they are totally aligned with similarly superior results obtained by the MT systems developed in the QTLeap project when they entered open competition in the WMT2016 shared task [Bojar et al., 2016].

HF company is thus quite optimistic that inclusion of MT will help to extend their business strategy in the future. The analysis of cost reduction reveals that the application of a deep linguistic-based system leads to important cost reductions in comparison with baseline statistical-based systems.

From the research perspective, we are convinced that the mix of evaluation strategies in QTLeap starting from automatic measures like BLEU via manual error annotation (reported in detail in Deliverable D2.11) up to the user evaluation reported in this Deliv-

erable is the right way to go and we hope that it can inspire other researchers in machine translation and related areas to take use-case evaluation more serious and integrate it into future research activities.

Taking into account all these different angles from which the MT systems developed in the QTLeap project were assessed, they are confluent in pointing towards the conclusion that for a vast majority of the language pairs in the project, the baseline Pilot 0 has been significantly, and in some cases substantially, outperformed by the machine translation technology resorted to and enhanced in this project, thus indicating its superior performance potential.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation (WMT16). In Ondřej Bojar and et al ., editors, *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*, volume 2, pages 131–198, Stroudsburg, PA, USA, 2016. ACL. ISBN 978-1-945626-10-4.

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. Chimera–three heads for English-to-Czech translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 90–96, 2013.

Rosa Gaudio, Aljoscha Burchardt, and António Branco. Evaluating machine translation in a usage scenario. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1–8. European Language Resources Association (ELRA), 2016.

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.