Evaluating Machine Translation in a Usage Scenario

Rosa Del Gaudio, Aljoscha Burchardt and António Branco
Higher Functions – Sistemas Inteligentes; DFKI – LT Lab; University of Lisbon
Lisbon, Portugal; Berlin, Germany; Lisbon, Portugal
rosa.gaudio@pcmedic.pt; aljoscha.burchardt@dfki.de; antonio.branco@di.fc.ul.pt

Abstract

In this document we report on a user-scenario-based evaluation aiming at assessing the performance of machine translation (MT) systems in a real context of use. We describe a sequel of experiments that has been performed to estimate the usefulness of MT and to test if improvements of MT technology lead to better performance in the usage scenario. One goal is to find the best methodology for evaluating the eventual benefit of a machine translation system in an application. The evaluation is based on the QTLeap corpus, a novel multilingual language resource that was collected through a real-life support service via chat. It is composed of naturally occurring utterances produced by users while interacting with a human technician providing answers. The corpus is available in eight different languages: Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish.

Keywords: Machine Translation, Evaluation, Multilingual Corpus

1. Introduction

Extrinsic evaluation of MT, i.e. assessment of 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 paper joins together general MT research and industrially focused applications of MT. The goal is to find the best methodology for evaluating the eventual benefit of a machine translation system in a real-world application of the type considered here by resorting to a user-based scenario approach.

This evaluation is based on the integration of MT services in a helpdesk application developed by the company Higher Functions as part of its business. It has been performed within the QTLeap project,[1] which aims to investigate an articulated methodology for machine translation based on deep language engineering approaches.

This paper is organised as follows. Section 2. provides some background information on the technical support scenario the evaluation is embedded in. Section 3. describes the QTLeap Corpus. Section 4. reports on the different experiments that constitute the evaluation. Finally, Section 5. concludes the paper.

2. Background: Technical Support Scenario

The usage scenario adopted in our evaluation is based on a service developed by the company Higher Functions Lda. to support their clients. This service, named PCWizard, offers technical support by chat, through a call-centre. It tries to automate the process of answering simple and recurrent user requests for IT troubleshooting for both hardware and software.

This problem solving procedure has been made efficient by using a Question Answering (QA) application and repository that helps call-centre operators prepare replies for clients.

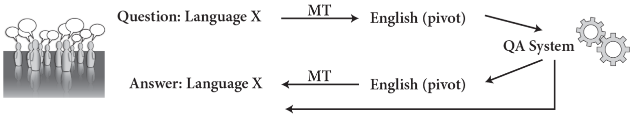


Figure 1: The workflow with the MT services

Using techniques based on natural language processing, each query for help is matched against a memory of previous questions and answers (QAs) 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 returned to the client. If the reply does not score over the threshold, the operator is presented with the list of possible answers delivered by the system and he can (a) pick the most appropriate reply, (b) modify one of the replies, or (c) write a completely new reply. In the last two cases, the new reply is used to further increase the QA memory.

Figure 1 shows the application workflow with embedded MT services. As the memory of previous question answering is in English, used as pivot language, there are two distinct places where MT services are used in the application: when the user request is translated to English in order to retrieve some previous interactions triggered by a similar query; and when the eventual answer that was retrieved is translated from English to the user's language.

3. The QTLeap Corpus

The QTLeap corpus is a novel multilingual language resource aiming at supporting multiple purposes in multilingual or cross-lingual applications such as evaluating the usefulness and quality of Machine Translation (MT) as described in the next sections of this paper. The corpus was created by the QTLeap project in order to extrinsically evaluate several MT engines in a task-based evaluation.

The QTLeap corpus is composed of question and answer pairs in the domain of computer and IT troubleshooting

[1] www.qtleap.eu

for both hardware and software. This corpus was collected through the PcWizard application, described in the previous Section.

While the original corpus was in Portuguese, professional translators have been contracted to produce a parallel corpus. The Portuguese corpus was first translated into English serving as the pivot language. The obtained English corpus was then translated into all the other languages addressed in the project. The actual corpus is available in eight different languages, namely Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish (Rosa Del Gaudio, 2015).

This kind of corpus is not very common, as most research is based on corpora using data sets composed by published texts, such as journal or books, or transcription of oral conversations. Furthermore, corpora with interrogatives are extremely rare, and most of them contain interrogatives that are artificially produced by manipulation over sentences that were originally declarative ones. The multilinguality of this corpus adds also to its importance.

In the last years, some corpora were collected composed by chat conversations over the internet. These corpora typically contain informal conversations about personal topics (Forsyth and Martell, 2007). Other corpora are more focused on technical issues such as the LINUX corpus (Elsner and Charniak, 2010), the IPHONE/PHYSICS/PYTHON corpus (Adams, 2008) and the Ubuntu chat corpus (Uthus and Aha, 2013).

These corpora differ from the one presented here as they include large amounts of utterances produce under social interaction, although the chats used as sources for these corpora were initially intended only for technical support. In all the referred corpora, the conversation threads involve several participants using an informal register. In almost all the cases (except for the Ubuntu corpus) the language addressed in these corpora is limited to English.

The corpus was collected selecting the data contained in the database of the PcWizard application, where all the interactions with the clients are saved. Interactions that better support the automatic QA module were selected. For this reason, only interactions composed by one question and the respective answer were included in the corpus.

3.1. Corpus Description

The QTLeap corpus is composed of 4000 question and answer pairs. It is characterized by short sentences, usually a request of help followed by an answer, and each conversation thread involves only two persons, the user and the operator. Some Portuguese examples with the English translations are provided below:

Question-PT Qual é a norma wireless mais recente?

Question-EN What is the latest wireless standard?

Answer-PT A norma mais recente é a norma N

Answer-EN The latest standard is the norm N

Question-PT Para ter internet em pontos afastados da casa recomendam PLC ou Repeater?

Question-EN In order to have internet all around the house, do you recommend PLC or Repeater?

Answer-PT A ligação através de PLC é mais estável e consoante a marca e modelo tem a possibilidade de também emitir rede sem fios. No entanto o repeater pode servir o propósito e ser uma solução mais económica.

Answer-EN The connection via PLC is more stable and depending on the brand and model it can also transmit wireless network. However, the repeater can reach the goal and it is a more economical solution.

This real-world corpus contains naturally occurring utterances and thus exhibits some characteristics of spoken language. The request for help is often a well-formed question or a declarative sentence reporting a problem, but in a relevant number of cases, the question is not grammatically correct, presenting problems with coordination, missing verbs, etc. In some cases, the request is composed only of a list of key words. This kind of utterance is representative of informal communication via chats. A somewhat more formal register characterizes the answers, as they are produced by well-trained operators and they need to be very precise and concise in order to clarify the user request and to not generate more doubts.

The professional translators were instructed to keep the informal register when translating from Portuguese to English, and from English to other languages, but to be as precise as possible regarding terminology.[2] We aimed to obtain a translation as closely as possible to the original language, but that still sounds natural to a native speaker of the target language.

On purpose, we have restrained from taking corrective actions to leave the character of the corpus intact. Yet, inspection has shown that even the answers contain several errors, such as, e.g., inconsistent use of terminology across answers, missing punctuation, or examples of "spoken style" such as enumeration of steps to be performed. For example, the answer below ends on three dots instead of a full-stop and a white-space is missing before the > symbol. This has been left unedited and has mostly been preserved also in the translations, e.g. into German.

Answer-PT Click File> Open...

Answer-DE Klicken Sie auf Datei> Öffnen ...

Table 1 provides some corpus statistics.

	Questions	Answers	Total
Tokens	50905	88536	139441
Sentences	4031	5919	9959
Tokens/Sentences	12.6	15	14
Sentences/Interactions	1	1.5	2.5

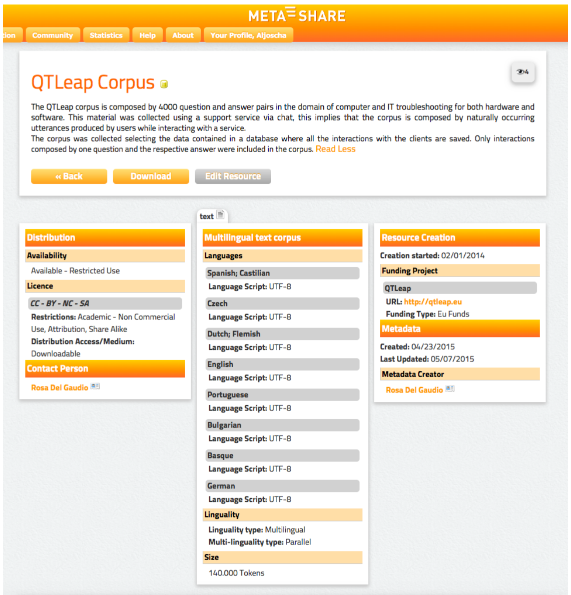
Table 1: Statistics (for English)

[2] No reference terminology was available for this newly created resource and it was beyond the aim and capacity of the project to create such a resource for all languages and the wide range of products/services covered.

3.2. License and distribution

The QTLeap corpus is available from META-SHARE[3] (see Figure 2). The type of license is Academic - Non Commercial Use, Attribution, Share Alike. It is parallel and includes all project languages Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish. The current version is v1.1.

For each language the corpus is composed by two plain text files, one listing all the questions and the other listing all the answers. The two files are aligned, this means that the query in the first line of the file containing all the questions corresponds to the reply in the first line of the file containing all the answers. This correspondence holds along the different languages.



QTLeap Corpus

The QTLeap corpus is composed by 4000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software. This material was collected using a support service via chat, this implies that the corpus is composed by naturally occurring utterances produced by users while interacting with a service.
The corpus was collected selecting the data contained in a database where all the interactions with the clients are saved. Only interactions composed by one question and the respective answer were included in the corpus. Read Less

Distribution
Availability
Available - Restricted Use
Licence
CC - BY - NC - SA
Restrictions: Academic - Non Commercial Use, Attribution, Share Alike
Distribution Access/Medium: Downloadable
Contact Person
Rosa Del Gaudio

Multilingual text corpus
Languages
Spanish; Castilian
Language Script: UTF-8
Czech
Language Script: UTF-8
Dutch; Flemish
Language Script: UTF-8
English
Language Script: UTF-8
Portuguese
Language Script: UTF-8
Bulgarian
Language Script: UTF-8
Basque
Language Script: UTF-8
German
Language Script: UTF-8
Linguality
Linguality type: Multilingual
Multi-linguality type: Parallel
Size
140.000 Tokens

Resource Creation
Creation started: 02/01/2014
Funding Project
QTLeap
URL: http://qtleap.eu
Funding Type: Eu Funds
Metadata
Created: 04/23/2015
Last Updated: 05/07/2015
Metadata Creator
Rosa Del Gaudio

Figure 2: The QTLeap Corpus on META-SHARE.

4. Experiments and Results

In general, 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 helpdesk in a multilingual environment. The main goals are to i) assess the eventual benefit of the MT services on the application, ii) find out to what extent the inclusion of MT can generate business opportunities, iii) set a baseline that makes it possible to see if future improvements of the MT technology lead to better performance in the usage scenario, and iv) taking into account the lessons learned with this exercise in view of coming up with a sound and viable methodology for extrinsic evaluation of MT.

All the four experiments reported below were carried out in a controlled setting in order to avoid dealing with different variables that interfere with the real objective of this evaluation. All the seven languages supported by the QTLeap corpus were tested.

The first experiment focuses on evaluating how the translation affects the answer retrieval component of the question and answer (QA) algorithm.

The remaining experiments focuses on outbound translation to evaluate to what extent it delivers a clear and understandable answer to final customers without the intervention of a human operator.

For Experiment 2 – 4, evaluators were recruited by project partners on a voluntary basis. Where possible, expert users were avoided as evaluators in order to simulate the typical usage of the PCWizard application by average laypersons to which it is directed. In fact, the volunteer evaluators generally presented a low or medium knowledge of the specific subject in question (around 80% of the times). This level of knowledge is typical of real customers of the PcMedic Wizard application and makes the results of the evaluation extensible to the real situation. In these three experiments, 100 question/answer-pairs have been evaluated, each by at least two human volunteers.

4.1. Experiment 1: Estimating the impact on the QA algorithm

The main idea of this experiment is to compare the results obtained when an original English question triggers the QA system with the results obtained when the questions are the result of the (intermediary) translation by the MT engine from a different language. That is, when English is acting as the pivot language for accessing the information encoded in the QA repository. This experiment was carried out automatically, without the intervention of human evaluators. A sample of the QTLeap corpus composed of 1000 interactions was selected for this experiment.

4.1.1. Setup

The QA system produces a list of candidate answers with a confidence score ranging between 1 and 100. 100 means that the QA search module is quite sure that the answer is correct for a given question. This score represents the confidence of the algorithm in monolingual answer retrieval. If the score is above 95 the answer is directly presented to the final user without the intervention of the human operator. When the score lies between 75 and 94, the answer is sent to the operator that can accept or modify it before sending it to the final user. The precise scoring algorithm is kept internal by HF company.

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 actual system.

Translations from the other project languages into English are produced using Moses SMT engines.[4]

4.1.2. Results

Table 2 shows the percentage of how many questions the QA algorithm is able to find a candidate answer for within a certain confidence score interval.

In general, when no translation is used, the English QA system can automatically answer a question without human intervention in 60% of the cases. In 34% it provides help for

[3] http://metashare.metanet4u.eu/go2/qtleapcorpus

[4] More information on the Moses engines can be found in the public project Deliverable D2.2: `http://qtleap.eu/wp-content/uploads/2015/06/QTLEAP-2015-D2.2.pdf`.

Score	EN	EU	BG	CS	NL	DE	PT	ES	Non-EN
>=95	60.2%	26.4%	19.7%	25.9%	23.1%	27.0%	20.7%	30.6%	24.8%
75-94	34.0%	26.4%	20.1%	23.8%	23.8%	24.9%	20.4%	32.8%	24.6%
50-74	5.2%	36.9%	41.3%	36.4%	39.5%	37.5%	43.0%	28.9%	37.6%
25-49	0.4%	9.0%	17.3%	12.4%	12.6%	9.6%	14.3%	6.4%	11.7%
1-24	0.2%	1.2%	1.5%	1.5%	1.1%	0.9%	1.5%	1.3%	1.3%

Table 2: Percentage of the answers delivered by QA System and their scores (recall)

	EN	EU	BG	CS	NL	DE	PT	ES	Non-EN
First	100%	76.4%	72.5%	77.9%	77.6%	82.1%	72.8%	86.0%	77.9%
Score	*91,8*	*78*	*71,6*	*75,5*	*74*	*76,7*	*71,8*	*80,2*	*75,4*
First two	100%	87.0%	85.0%	87.2%	87.6%	91.4%	84.3%	92.8%	87.9%
Score	*91,8*	*75,5*	*68,8*	*73,5*	*71,8*	*74,5*	*69,4*	*78,8*	*73,2*
First three	100%	90.1%	88.3%	91.1%	90.3%	93.8%	87.8%	94.5%	90.1%
Score	*91,8*	*74,7*	*67,8*	*72,3*	*70,9*	*73,7*	*68,5*	*78,4*	*72,3*

Table 3: Percentage of answers delivered as first candidates for both English and target language (precision)

operators in finding the right answer by supporting them with retrieved candidate answers. Only in about 6% of the cases the operator is left without any help. When the translation services are used, the answers scoring 95 or more drop considerably, from 19% of Bulgarian to 31% of Spanish. The answers with a score above 75 represent about half of the cases.

Table 3 shows the percentage of cases in which the first answer of the gold standard appears in the list of the answers obtained using the translated question, particularly if it appears in the first place, in the first two or in the first three places. The mean score for this answer is also presented. For example, in 76.4% of Basque cases the first retrieved answer had the same ID as the best scoring English answer, and on average this answer gets a score of 78.

Overall, these data indicate that on average in almost 78% of the cases when MT is used the first answer suggested by the QA system corresponds to the first suggestion of the gold standard, which is a very promising result.

Qualitative inspection has shown that the scores will probably need to be adjusted for individual languages and the threshold for sufficiently good answers might well be lower than 95. Therefore, the tendencies observed in this evaluation will probably lead to even greater positive effects in reality.

4.2. Experiment 2: Estimating probability of operator calls

Based on the user scenario at stake, a test setup and metric was elaborated that attempts to determine the probability of the users making a phone call to get a satisfactory answer to their questions in case the interaction via chat was eventually felt as not satisfactory by the user. The less phone calls the better from a commercial perspective. And the less phone calls the better in terms of the contribution of the MT engine to support the multilingual deploying of the helpdesk service via chat.

The MT engines used in this experiment were Moses SMT engines as in Experiment 1.

4.2.1. Setup

At a basic level, this evaluation exposes the human evaluator first to the machine translated (MT) answer and then to the reference answer. In this way, the subject evaluates the MT answer first on its own and then with respect to the reference.

Using a web interface, a question is presented to the evaluator in the target language followed by the automatically translated answer (A). In this step the subject assesses the usefulness of this answer.

After answering, the evaluator is presented again with the question, the MT answer (A), and this time also with the reference answer (B). The subject is now asked to compare answers A and B, taking into account that the second answer B is giving the correct information.

In this experiment, all the question/answer pairs were evaluated at least by 3 volunteers for each of the seven languages, with a global average of 3.3 evaluators per pair.

4.2.2. Results

Table 4 shows the evaluation results when the evaluator is asked to assess the usefulness of the automatically translated answer. Based on these results, the quality of the response is very different across the languages.

Table 5 reports on the results when the evaluator was asked to compare the automatically translated answers (A) with the reference answer (B) giving the correct information.

When the reference answer is presented, different results are obtained compared to the previous table. In particular, the evaluations are more homogeneous among all the languages and among the three different options.

To interpret these results, we designed a metric that assesses the probability of operator calls. What it is relevant for this metric is the perception of the user about the correctness of the answer. This means that if the evaluator appreciation is that the automatically translated answer would clearly help to answer the question at stake, then the probability of asking for further help by picking the phone would very low. This would be the case especially if the answer, when compared to a reference answer, is judged as giving the right advice or being just some minor points wrong.

	EU	BG	CS	NL	DE	PT	ES	Avg.
It would clearly help me solve my problem / answer my question	30.7%	48.1%	49.5%	24.7%	37.3%	12.4%	65.3%	38.3%
It might help, but would require some thinking to understand it	47.7%	43.6%	35.2%	43.4%	41.4%	35.3%	26.3%	39.0%
It is not helpful / I don't understand it	21.7%	8.3%	15.3%	31.6%	21.3%	52.3%	8.3%	22.7%

Table 4: Assessment of the usefulness of the translated answers

	EU	BG	CS	NL	DE	PT	ES	Avg.
A gives the right advice	25.7%	35.0%	42.2%	25.6%	43.2%	22.9%	45.3%	34.3%
A gets minor points wrong	37.7%	44.3%	31.9%	35.9%	33.4%	23.2%	22.3%	32.7%
A gets important points wrong	36.7%	20.7%	25.9%	38.4%	23.4%	54.0%	32.3%	33.1%

Table 5: Assessment of the translated answer against the reference answer

MT answer judged alone	MT answer compared with reference answer	Probability
Solves my problem	Gets the right advice	low
Solves my problem	Gets minor points wrong	low
Would require some thinking to understand it	Gets the right advice	low
Would require some thinking to understand it	Gets minor points wrong	medium
Solves my problem	Gets important points wrong	high
Would require some thinking to understand it	Gets important points wrong	high
Is not helpful / I don't understand it	Gets the right advice	high
Is not helpful / I don't understand it	Gets minor points wrong	high
Is not helpful / I don't understand it	Gets important points wrong	high

Table 6: The metric with the probability of calling an operator

Probability	EU	BG	CS	NL	DE	PT	ES	Avg.
low	33.3%	47.4%	54.5%	30.4%	47.8%	21.5%	60.4%	42.2%
medium	28.1%	30.6%	17.9%	21.9%	22.0%	15.8%	7.0%	20.5%
high	37.0%	22.0%	27.5%	47.7%	30.1%	62.7%	32.7%	37.1%

Table 7: Aggregated results of the metric providing the probability of calling an operator

Table 6 shows the probability of calling an operator for each different possibility.
Finally, Table 7 combines the results of the previous tables and report on the probability of calling an operator.

4.3. Experiment 3: User ranking

In another experiment that follows the basic setup of Experiment 2, yet comparing three MT engines, we wanted to see how the user assessment relates to the difference between the engines (called Pilot 0, Pilot 1, and Pilot 2 or P0, P1, P2 for short)[5] performance in terms of BLEU scores.
In this evaluation, instead of rating the usefulness, we asked evaluators to rank the three different translations delivered by Pilot 0, Pilot 1 and Pilot 2, which is common practice in human MT evaluation, e.g., as performed in the WMT Shared Tasks.

4.3.1. Setup

In a web form, one question at a time is presented, followed by the manually translated answer and by the three answers generated by Pilot 0, Pilot 1 and Pilot 2 (anonymized as "A", "B" and "C", in random order, so the evaluation is blind). The subject is asked to read the reference answer and the three following answers. Then, being told to suppose that the reference answer is correct, the evaluator is asked to rank the three answers from best to worst. It is possible to assign the same rank to more than one answer. The precise instructions to the annotators were (presented in their mother tongue):

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

[5] While the Pilot 0 systems are the Moses SMT baselines used in experiment 2, the Pilot 1 engines are slightly "deeper" MT systems as documented in (Popel et al., 2015b) while Pilot 2 engines were MT systems extended by lexical semantic knowledge, see (Popel et al., 2015a).

	EU	**BG**	**CS**	**NL**	**DE**	**PT**	**ES**
P0	1.31	1.94	2.17	1.87	2.35	2.33	2.15
P1	2.27	2.09	2.22	2.03	2.39	2.42	2.36
P2	1.93	2.12	2.07	1.81	2.33	2.18	1.78

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

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

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

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

Table 10: Comparison between Pilot 2 and Pilot 0

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

Table 11: Comparison between Pilot 2 and Pilot 1

4.3.2. Results

Table 8 shows the average score obtained with the ranking, where 1 means best and 3 worst. This table offers a first insight in the performance of the different pilots.

For all the languages, Pilot 1 answers obtain a worse score than Pilot 0. For Basque, Pilot 2 shows an improvement over Pilot 1, but not over Pilot 0. For Bulgarian, the best results are obtained by Pilot 0, followed by Pilot 1. For the remaining five languages, Pilot 2 outperforms both Pilot 1 and Pilot 0.

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

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

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

Figures 3 and 4 provide a graphical representation of Tables 10 and 11, respectively. The languages (vertical bars) in the figures are sorted by the *better ignoring ties* scores, which are plotted as a dark blue square. We can see that for four languages, Pilot 2 is better than the respective Pilot 0 (the *better ignoring ties* score is higher than 50%). Also, all languages except for Basque have Pilot 2 at least as good as the respective Pilot 0 (the yellow bar reaches over 50%).

From Figure 4 we can see that for all languages except for Bulgarian, Pilot 2 is better than Pilot 1.

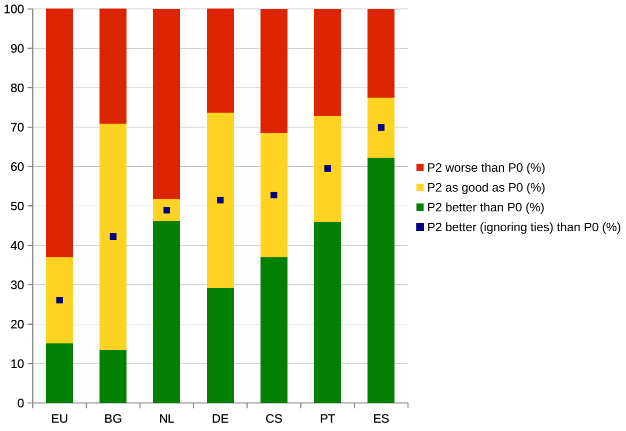


Figure 3: Comparison of Pilot 2 and Pilot 0, breakdown of the human evaluation. In each bar, top, midlle and bottom segments represent, respectively, “worse than”, “as good as” and “better than”.

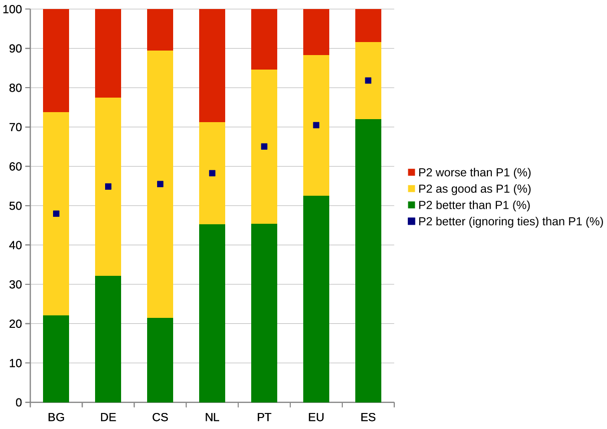


Figure 4: Comparison of Pilot 2 and Pilot 1, breakdown of the human evaluation. In each bar, top, middle and bottom segments represent, respectively, “worse than”, “as good as” and “better than”.

4.4. Experiment 4: Comparing user preferences with BLEU differences

We finally compared this human extrinsic evaluation data with the automatic intrinsic performance measure BLEU. Table 12 shows the BLEU scores of the three pilots and the difference between Pilot 2 and the other two Pilots.

Figures 5 and 6 present the BLEU difference (dark red bars) in relation to the difference between Pilots according to the human extrinsic evaluation (violet bars).

For this purpose, we scaled the *better ignoring ties* score (defined in Section 4.3.2.) to the same range as BLEU difference, that is $\langle -100; +100 \rangle$, which boils down to %(P2 better than PX) - %(P2 worse than PX), where PX means Pilot 0 (in Figure 5) or Pilot 1 (in Figure 6). The languages (bars) in Figures 5 and 6 are sorted according to this human evaluation (that is, in the same order as in Figures 3 and 4, respectively.)

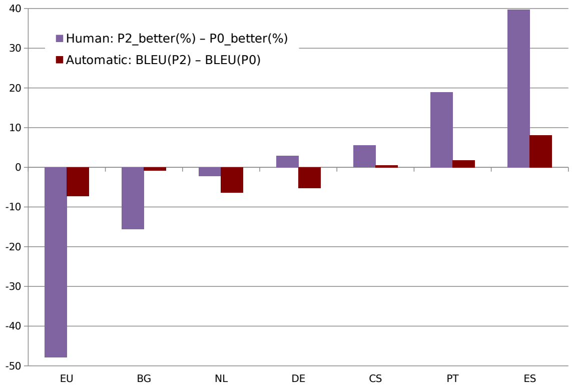


Figure 5: Comparison of user evaluation results and BLEU scores for Pilot 2 and Pilot 0. In each pair of bars, the left and the right bars stand, respectively, for “human” and “automatic” evaluation.

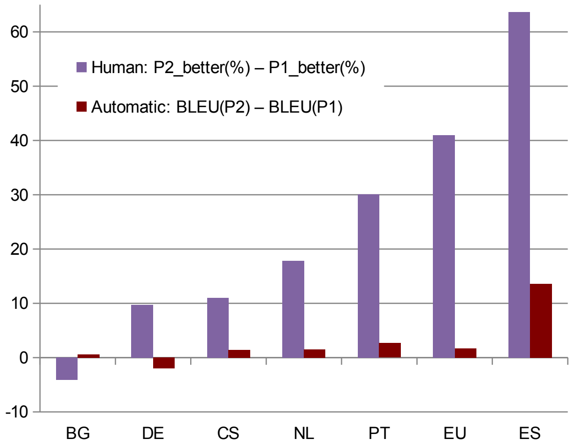


Figure 6: Comparison of user evaluation results and BLEU scores for Pilot 2 and Pilot 1. In each pair of bars, the left and the right bars stand, respectively, for “human” and “automatic” evaluation.

It is interesting that BLEU differences almost always agree with the user ratings on the comparison of two systems. There are just three exception: German Pilot 2 vs. Pilot 0 (see Figure 5), and German and Bulgarian Pilot 2 vs. Pilot 1 (see Figure 6). For German, the user ratings are generally more in favor of Pilot 2 (unlike BLEU).

5. Conclusions

In this paper we have presented an innovative method to evaluate MT systems, taking into consideration real user context. We have shown a sequel of experiments that focuses on different aspects to be evaluated in different MT setups.

This extrinsic evaluation of Machine Translation systems created in QTLeap has compared the performance of three different MT systems for each one of the seven project languages.

The MT systems have been tested by volunteer subjects in a usage scenario of project partner HF, namely a chat-based PC helpdesk scenario (PcWizard).

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

Table 12: Comparison between all Pilots in terms of BLEU on QTLeap Corpus (Batch3[6]).

We hope that this work can serve as an inspiration for other researchers in this area that helps to establish task-based evaluation in the growing mix of MT evaluation strategies. We also presented the QTLeap Corpus, a novel multilingual language resource aiming at supporting multiple purposes in multilingual or cross-lingual applications. The corpus is composed of 4000 question and answer pairs from the IT helpdesk domain. This real-world corpus contains naturally occurring utterances and thus exhibits some characteristics of spoken language. It is unique given the fact that it is a data set with parallel utterances in eight different languages (Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish), from four different language families (Basque, Germanic, Romance and Slavic).

Acknowledgements

This work has received support by the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches".

6. Bibliographical References

Adams, P. H. (2008). Conversation Thread Extraction and Topic Detection in Text-Based Chat. Master's thesis, Naval Postgraduate School, Monterey, California.

Elsner, M. and Charniak, E. (2010). Disentangling chat. *Comput. Linguist.*, 36(3):389–409, September.

Forsyth, E. and Martell, C. (2007). Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 19–26, Sept.

Popel, M., Dušek, O., Branco, A., Gomes, L., Rodrigues, J., Silva, J., Avramidis, E., Burchardt, A., Lommel, A., Aranberri, N., Labaka, G., van Noord, G., Gaudio, R. D., Novák, M., Rosa, R., Hlaváčová, J., Hajič, J., Todorova, V., and Popov, A. (2015a). Report on the second mt pilot and its evaluation. *QTLeap Project*, Deliverable D2.8(http://qtleap.eu/wp-content/uploads/2015/11/QTLEAP-2015-D2.81.pdf).

Popel, M., Hlaváčová, J., Bojar, O., Dušek, O., Branco, A., Gomes, L., Rodrigues, J., Silva, J., Querido, A., Rendeiro, N., Campos, M., Amaral, D., Avramidis, E., Burchardt, A., Popovic, M., Lommel, A., Simova, I., Aranberri, N., Labaka, G., van Noord, G., Gaudio, R. D., Novák, M., Rosa, R., Tamchyna, A., and Hajič, J. (2015b). Report on the first mt pilot and its evaluation. *QTLeap Project*, Deliverable D2.4(http://qtleap.eu/wp-content/uploads/2015/04/QTLEAP-2015-D2.4.pdf).

Uthus, D. C. and Aha, D. W. (2013). The ubuntu chat corpus for multiparticipant chat analysis. In *Analyzing Microtext, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013.*

7. Language Resource References

Rosa Del Gaudio. (2015). *QTLeap Corpus*. QTLeap Project, distributed via META-SHARE.