Evaluating a Machine Translation System in a Technical Support Scenario

Rosa Del Gaudio*, **Aljoscha Burchardt and Arle Lommel**
* Higher Functions – Sistemas Inteligentes
Lisbon, Portugal
rosa.gaudio@pcmedic.pt
German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab - Berlin, Germany
aljoscha.burchardt@dfki.de
arle.lommel@dfki.de

Abstract

In this document we report on a user scenario based evaluation aiming at assessing the performance of a machine translation (MT) system in a real context of use. This extrinsic evaluation exemplifies a framework that makes it possible to estimate MT performance and to verify if improvements of MT technology lead to better performance in a real usage scenario. We report on the evaluation of Moses baselines for several languages in a cross-lingual IT helpdesk scenario.

1 Introduction

Extrinsic evaluation of MT, i.e., assessment of MT quality 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 research and industrial applications. The focus is to find the best procedure for evaluating a machine translation system in a real-world application using a user-based scenario methodology.

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 (see Section 3) to make it cross-lingual. 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 and evaluates several different MT approaches in a usage scenario.

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 system of the helpdesk in a multilingual environment. The main goals are to i) assess the impact of the MT services on the application, ii) find out to what extent the inclusion of MT can generate business opportunities, and 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.

In order to reach this objective, the evaluation was split in two distinctive parts. The first part focuses on evaluating how the translation affects the answer retrieval component of the question and answer (QA) algorithm. The second part focuses on outbound translation to evaluate to what extent it delivers a clear and understandable answer to final customers without the intervention of a human operator. In this paper we report on the second part of the evaluation covering seven different languages: Basque, Bulgarian, Czech, Dutch, German, Portuguese and Spanish.

Section 2 reports on the state of the art, while Section 3 describes the real user scenario. Section 4 explains in details how the evaluation was carried out. Section 5 presents the results for each of the seven languages. Finally we draw some conclusions in Section 6.

[1] www.qtleap.eu

Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015), pages 39–47,
Praha, Czech Republic, 3–4 September 2015.

2 State of the art

Previous work includes extrinsic evaluation of machine translation through several MT applications: cross-lingual patent retrieval, cross-lingual sentiment classification, collaborative work via idea exchange, speech-to-speech translation, and dialogue.

The Patent Translation Task at the Seventh NTCIR Workshop employed search topics for cross-lingual patent retrieval, which was used to evaluate the contribution of machine translation for retrieving patent documents across languages (Fujii et al., 2008). They also analysed the relationship between the accuracy of MT and its effects on retrieval accuracy (Fujii et al., 2009), which comes closest to the evaluation of answer retrieval in our scenario.

Duh et al. (2011) investigated the effect of Machine Translation on Cross-lingual Sentiment classification and suggested improvements to the adaptation problems that have been identified. Yamashita and Ishida (2006) started research on collaborative work using machine translation. Similarly, Wang et al. (2013) evaluated MT through idea exchange: in this scenario, pairs of one English and one Chinese speaker performed brainstorming tasks assisted by MT, which helped the non-native English speakers produce ideas; nevertheless comprehension problems were identified with MT output.

In the early years of NLP, the Verbmobil project (Jekat and Hahn, 2000) performed end-to-end Machine-Translation as part of a longer pipeline with several modules, and evaluation of MT via speech-to-speech translation has been conducted in the frame of a yearly shared task (e.g., Cellotolo *et al.* (2013)). In another example on dialogue systems, Schneider *et al.* (2010) employed a "Wizard of Oz" technique in order to assess the quality of translations in the context of a dialogue application. A human operator (the "wizard") who is not visible to the user, takes the role of the system. In that scenario, German speakers have to find a good offer on Internet connections in Ireland. The extrinsic evaluation measuring elapsed time, shows different results to the intrinsic error-specific MT evaluation. The questionnaire we use in our evaluation is based on the one used in this task.

3 Tech support scenario

The scenario used in our evaluation is based on a real service developed by the Portuguese company Higher Functions to support their clients. This service, named PcWizard, offers technical support by chat. Usually technical support can be divided into three levels based on the difficulty of the request: first-level, second-level, and third-level. Most of the first-level inquiries are straightforward and simple, and can be easily handled. Literature has shown that the majority of user requests can be answered by the front-line 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 all requests and are responsible for long waiting times, leading to user dissatisfaction. The PcWizard 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 centre. This process of problem solving can be made efficient by using a Question Answering (QA) application that helps call centre 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 (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 the embedded MT services. As the memory of previous question answering is in English, 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. This translation is used by the QA search algorithm for retrieving a possible answer.

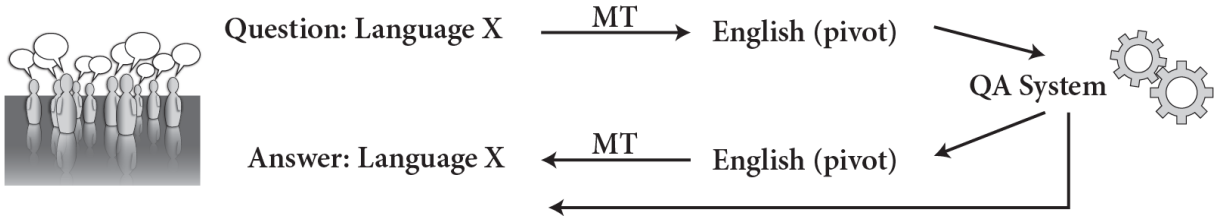

Figure 1: The workflow with the MT services

Once an answer is found in English, MT services are used to translate the answer back to the users original language. This means that the MT services interact with the system in two different moments 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. (In this case all the translations are done to or from English.)

4 The experimental settings

This evaluation was carried out in a controlled setting in order to avoid dealing with different variables that interfere with the real objective of this evaluation, such as having a relatively small multilingual database and no previous data on a multilingual scenario. Furthermore a direct field test would lead to the problem that the questions would differ between evaluations and complicate comparison of the results. For these reasons 100 question/answer pairs from the corpus were selected and volunteers were recruited for this in-vitro experiment. Where possible, IT experts were avoided as evaluators in order to simulate the typical user of the PcWizard application.

4.1 The corpus

A corpus was collected to develop and evaluate different MT systems. This corpus is composed of 4000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software. As this corpus was collected using the PcWizard application, it is composed of naturally occurring utterances produced by users while interacting with the service. The corpus was collected by selecting the data contained in the database of the application, in which all client interactions are saved. For this corpus, only interactions composed of one question and a respective answer were considered.

The corpus consists of short sentences (usually a request of help) followed by an answer, and each conversation thread involves only two persons (the user and the operator). The request for help is often a well-formed question or a declarative sentence reporting a problem, but in a significant number of cases, the question is not grammatically correct, presenting problems with coordination, missing verbs, etc. In some cases, the request is composed of a list of key words. This kind of utterance is representative of informal communication via chats. On the other hand, a 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 avoid generating more confusion.

The corpus, available for Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish, can be downloaded from the META-SHARE portal [2] under the name “QTLeap Corpus”.

4.2 Evaluation workflow

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 and then he/she is asked to provide a self-estimation of his/her knowledge level (high, medium, or low) on the subject involved in the question.

[2] http://metashare.metanet4u.eu/

	A	SA	N	SD	D
I have serious problems in understanding this answer	-	-		+	+
These sentences are fluent	+	+		-	-
There are awkward words and expressions	-	-		+	+
I would rate the sentences as comprehensible	+	+		-	-
Some words appear in a strange order	-	-		+	+
The instructions/information in the answer are not clear	-	-		+	+
I would consider using a similar system for technical support in a similar context	+	+		-	-

Table 1: List of the statements used in the final questionnaire

Then the same question is presented, followed by the automatically translated answer (A). In this step the subject assesses on the usefulness of this answer, according to the following options:

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

After answering, the evaluator is presented again with the question, the MT answer (A), and the reference answer (B). This time the subject is asked to compare answers A and B. Taking into account that the second answer B is giving the correct information, he/she is asked to re-evaluate the first answer A, selecting one of the following options:

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

Finally, evaluators are asked to give a closer look at the automatically translated answer and provide a more fine-grained evaluation on seven different aspects using a questionnaire with answers based on a 5-point Likert scale: agree (A), slightly agree (SA), neither agree nor disagree (N), slightly disagree (SD), disagree (D). At this point evaluators have also the possibility to leave a comment on the interaction evaluated.

The statements used for this questionnaire were developed using the questionnaire presented in (Schneider et al., 2010) for evaluating an MT dialogue system as a starting point. Following the literature, the statements in the questionnaire were designed in order to balance the number of negative and positive statements to avoid getting the same judgment and to force evaluators to read each statement carefully.

Table 1 shows the list of the statements used in the questionnaire. The plus and minus symbols represent the value of the statement. A positive judgment is represented by the plus, a negative by minus. For instance, if the evaluator agrees with the first statement, it means that the sentence presents some kind of problem, so it is negative for the performance of the system. For the second statement (which presents a positive judgment) the situation is inverted: if the evaluator agrees, it means that, for that specific aspect, the answer present a positive score.

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.

5 Results

To clarify the framework, this section presents the results of evaluating Moses baselines for the project languages mentioned above. Is is important to note that it is not our goal to compare performance between languages, even if the presentation of results might raise this expectation. The Moses systems

that have been set up for the different languages have been trained on different general and domain corpora depending on availability of resources. Table 2 shows the evaluation results when the evaluator is asked to assess on the usefulness of the automatically translated answer. Based on these results, the quality of the response is very different across the languages.

	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 2: Assessment of the usefulness of the translated answers

Bulgarian and Spanish received the best evaluation with only 8.3% of answers judged as not helpful/not understandable, versus 52.3% for Portuguese. Czech also demonstrated good performance, with almost 50% of the answers considered clearly helpful in answering the question.

Table 3 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, very different results were obtained compared to the previous table. In particular, the evaluations are more homogeneous among all the languages and among the three different options.

It is interesting to note that the positive evaluation obtained when only the MT answer is presented decreases for four of the seven languages (Basque, Bulgarian, Czech and Spanish), but increases for the other three (Dutch, German, Portuguese). Subjects using Dutch and Portuguese were the ones providing the worst evaluation of MT answers.

Based on this scenario, a metric was elaborated. This metric attempts to determine the probability of final users making a phone call to get a satisfactory answer to their questions. 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 checked that the automatically translated answer would "clearly help to solve my problem/answer the question" the probability of asking for further help 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 just some minor points wrong.

Cases when an evaluator thinks that the translated answer would require some thinking to understand it and gets important points wrong are rather different: in this case the probability of calling an operator would be higher.

Table 4 shows the probability of calling an operator for each different possibility. The results for each language are presented in Table 5.

In order to draw some considerations, the aggregates results are presented in Table 6.

As noticed in the previous tables, there is a high degree of variance between the different languages. For example, Spanish or Czech present a much smaller probability of users calling an operator than do Portuguese and Dutch.

	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 3: Assessment of the translated answer against the reference answer

	MT answer	Reference Answer	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 4: The metric with the probability of calling an operator

	Probability	EU	BG	CS	NL	DE	PT	ES	Avg.
A	low	20.8%	28.6%	34.9%	14.4%	29.0%	8.8%	39.7%	25.2%
B	low	7.9%	14.8%	12.6%	8.8%	7.2%	2.5%	15.0%	9.8%
C	low	4.6%	4.0%	7.0%	7.2%	11.6%	10.2%	5.7%	7.2%
D	medium	28.1%	30.6%	17.9%	21.9%	22.0%	15.8%	7.0%	20.5%
E	high	1.7%	1.3%	2.0%	1.6%	1.4%	1.1%	10.7%	2.8%
F	high	14.9%	11.7%	10.3%	14.4%	7.8%	9.3%	13.7%	11.76%
G	high	0.3%	0.0%	0.3%	4.1%	3.2%	4.0%	0.0%	1.7%
H	high	1.3%	1.5%	1.3%	5.3%	3.8%	4.8%	0.3%	2.6%
I	high	20.5%	7.5%	13.6%	22.3%	13.9%	43.5%	8.0%	18.5%

Table 5: Results of the metric considering each case

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 6: Aggregated results of the metric

The following graphics report on the results obtained with the final questionnaire where the MT answers were evaluated on seven different aspects: understanding, fluency, awkwardness, word order, clarity, use of this type of system.

The agree/disagree evaluation are normalised into positive/negative judgments and then calculated as a weighted average where the slightly positive/negative cases get a lower weight of .5 and neutral values are simply ignored: (positive - negative) + 0.5*(slightly positive - slightly negative).

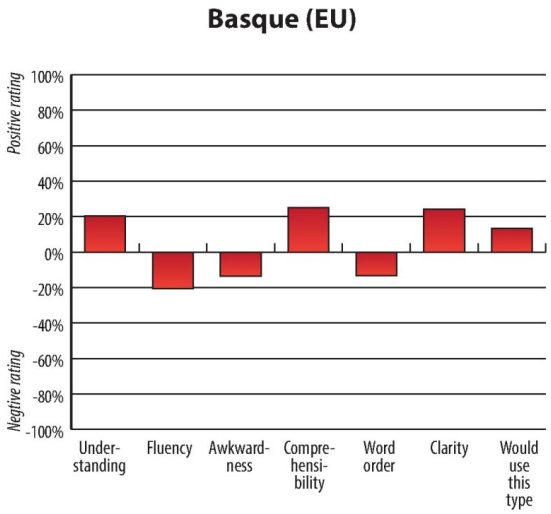

Figure 2: Basque

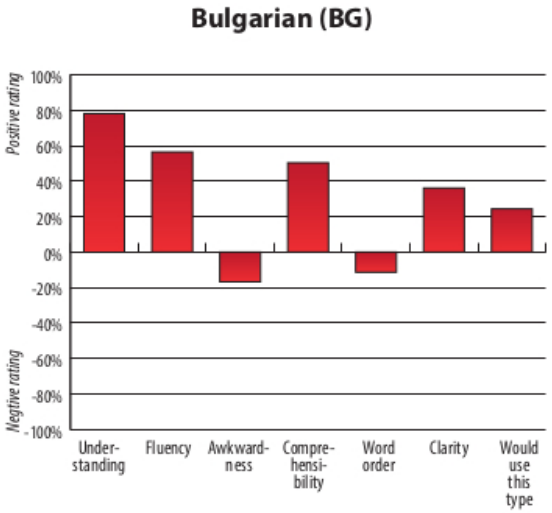

Figure 3: Bulgarian

As show in Figure 2, the Basque speaking subjects provide positive evaluation in four out of seven statements. In particular evaluators agree in the 54.5% of the cases that they do not have serious problems in understanding the answer, that it was comprehensible (56.5%), and the instructions were clear (54.5%), and they would consider using a similar system (47.8%). The problems come up with the lack of fluency of the sentences (51.5%), the presence of awkward expressions (58.6%) and the order of the words (61.8%).

For Bulgarian (Figure 3) the outcome was more positive than for Basque. The positive dimensions go from four to five. The sentences are also considered fluent in 78.3% of the cases. In general the positive dimension obtains higher values. For example, in 89.2% of the cases the evaluators have no problems in understanding the answer. The problem again is the presence of awkward expressions (58.6%) and the order of the words (54.5%).

For Czech, Figure 4, all the dimensions are positives with the exception of the fluency of the sentences where positive and negative judgments present almost the same weight (41.7% and 42.7% respectively).

Figure 5 shows a very opposite evaluation for Dutch speaking subjects. All the statements get a negative evaluation.

German-speaking evaluators provided positive evaluation on three dimensions: the understandability of the answer (50.6%), the clarity of instructions (49.1%) and use of the system in a similar situation (34.6%).

Portuguese speaking evaluators, similarly to Dutch, provide negative evaluation on all the seven statements.

Finally, Spanish-speaking subjects evaluated six of the seven dimensions positively. The only problem is given by the presence of awkward words and expressions reported in 63.1% of the evaluations.

6 Conclusions

In this paper we presented an innovative method to evaluate MT systems, taking into consideration real user context.

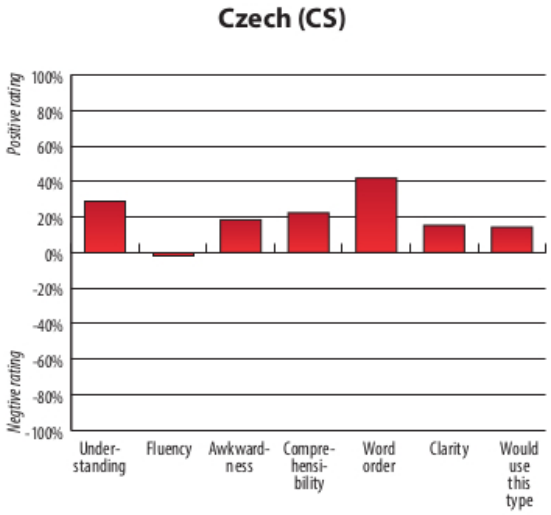

Figure 4: Czech

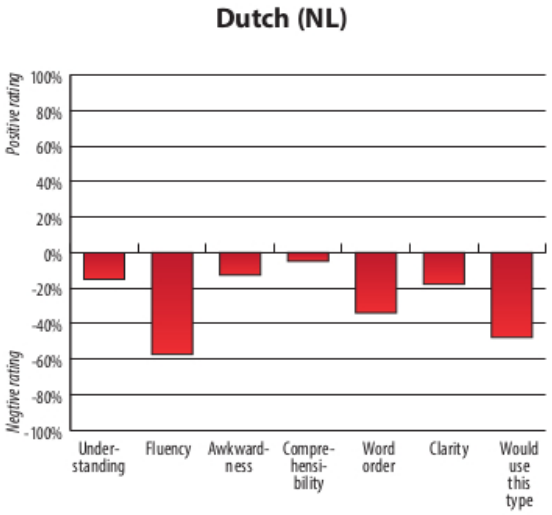

Figure 5: Dutch

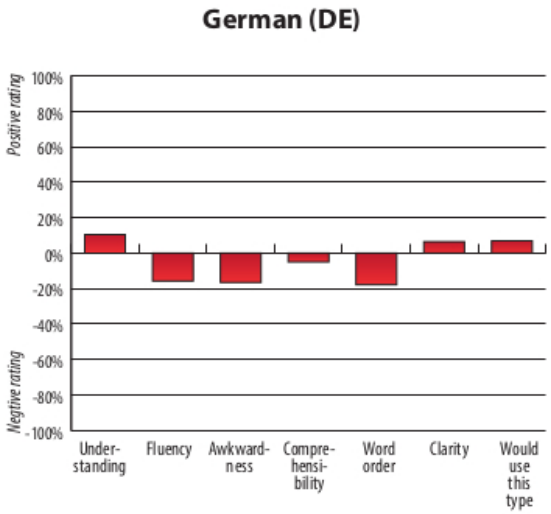

Figure 6: German

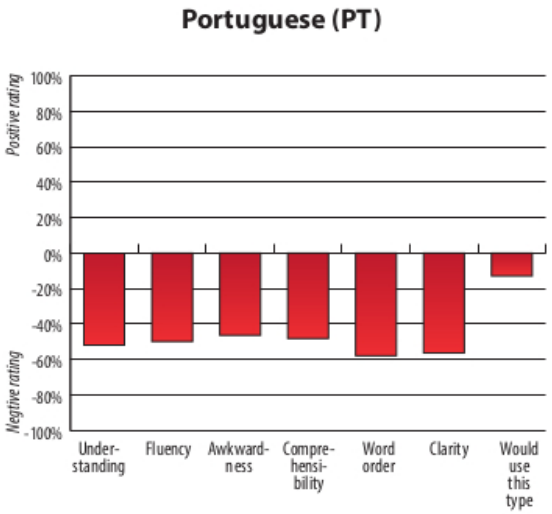

Figure 7: Portuguese

With this evaluation we show that although the translations of answers presented to the real user were produced by baseline systems, results are promising from a business perspective and could result in a real-world reduction in service calls. Even if there are many flaws in the translations, a considerable part of the test users would use a system like this again and the approximated chance of calling an operator is lower than expected (even allowing for the fact that the numbers are approximations).

The results reported in this paper provide the basis for the extrinsic evaluation of the impact of MT system to the QA system where it was embedded.

Acknowledgements

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

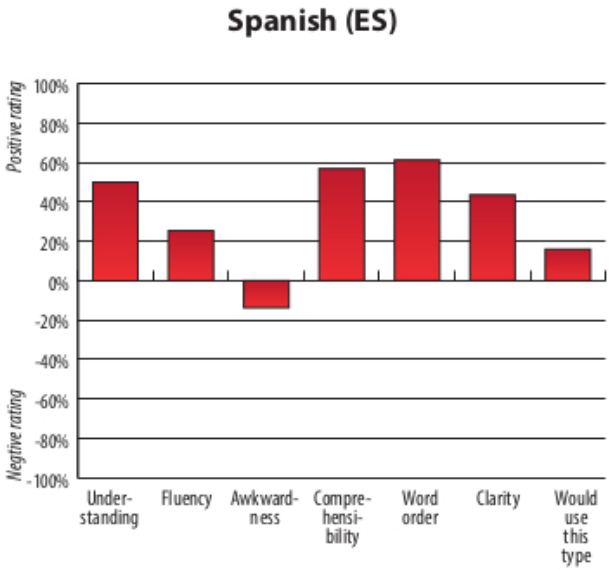

Figure 8: Spanish

References

Mauro Cettolo, Jan Niehues, Sebastian Stker, and Luisa Bentivogli andMarcello Federico. 2013. Report on the 10th iwslt evaluation campaign. In *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013)*, pages 29–38.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 429–433, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. 2008. Overview of the patent translation task at the ntcir-7 workshop. In *In Proceedings of the 7th NTCIR Workshop Meeting*, pages 349–400.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2009. Evaluating effects of machine translation accuracy on cross-lingual patent retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 674–675, New York, NY, USA. ACM.

SusanneJ. Jekat and Walther Hahn. 2000. Multilingual verbmobil-dialogs: Experiments, data collection and data analysis. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 575–582. Springer Berlin Heidelberg.

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

Anne Schneider, Ielka van der Sluis, and Saturnino Luz. 2010. Comparing intrinsic and extrinsic evaluation of mt output in a dialogue system. In *IWSLT*, pages 329–336.

Hao-Chuan Wang, Susan Fussell, and Dan Cosley. 2013. Machine translation vs. common language: Effects on idea exchange in cross-lingual groups. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 935–944, New York, NY, USA. ACM.

Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, pages 515–524, New York, NY, USA. ACM.