qtleap
quality
translation
by deep
language
engineering
approaches

REPORT ON THE EXTRINSIC EVALUATION METRICS

DELIVERABLE D3.3

VERSION 1.0 | 2014 JUNE 30

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	author	organisation	description
0	2014 June 13	Aljoscha Burchardt	DFKI	First draft including input from all partners
1	2014 June 18	Eleftherios Avramidis, António Branco, Aljoscha Burchardt, Rosa Del Gaudio	DFKI, FCUL, HF	First stable version
1	2014 June 30	Jan Haijč	CUNI	Internal review

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 EXTRINSIC EVALUATION METRICS

DOCUMENT QTLEAP-2014-D3.3
EC FP7 PROJECT #610516

DELIVERABLE D3.3

completion

FINAL

status

SUBMITTED

dissemination level

PUBLIC

responsible

ALJOSCHA BURCHARDT (WP3 COORDINATOR)

reviewer

JAN HAJIČ

contributing partners

FCUL, DFKI, HF

authors

ALJOSCHA BURCHARDT, ANTÓNIO BRANCO, ROSA DEL GAUDIO

1. Introduction

The extrinsic evaluation described in this Deliverable complements the intrinsic evaluations of deep MT (Task 2.5) by evaluating the MT Pilots in terms of their impact on the performance of the QA system of the helpdesk in which they will be embedded (see Task 3.3).

Main driver of the extrinsic evaluation is the QTLeap (industry) partner HF who is running the helpdesk as part of their business. The methodology outlined below has been developed throughout a number of intensive discussions and exchange of drafts and ideas via email among the involved partners. From the very beginning, HF had uttered the wish to proceed in a practical fashion where pragmatic solutions and flexibility ensure that HF can use the evaluation results for their business development. In this spirit, this Deliverable is to be understood as a living document. The evaluation methods described below will be tested on the MT baselines and be revised after a critical review.

1.1. Main challenges

From the perspective of the enterprise, a QA system variant B improves over a variant A if there are -- on the average -- fewer interventions of human operators necessary, irrespective of whether an MT system is involved or not. The main challenge in the design of extrinsic evaluation methodology was to factor out sensible checkpoints in the QA workflow that make it possible to assess (only) the contribution of the MT systems to the number of human interventions. Another challenge was that as HF did not offer multilingual services in the past so that there is no natural baseline available.

While the one extreme (measuring human interactions in an MT-based scenario as in a monolingual scenario) would bare the risk of assessing the overall QA service quality, but not the particular contribution of MT, the other extreme (measuring accuracy and fluency of the MT output and deducing QA performance from it) would be too much of an intrinsic and implicit evaluation. In the suggested methodology below, the partners have tried to find a good balance between the two extremes.

The evaluation has been designed as to on the one hand help HF in business development and at the same time provide valuable input to the MT development by the research partners in the project and help compare the performance of the MT pilots and correlate the extrinsic performance values with intrinsic MT quality measures.

1.2. State of the Art

Extrinsic evaluation of MT has not (yet) established itself as a major research topic. Reasons may include the prevalent focus of MT research on gisting translation of newspaper texts, which does not lend itself too much to task-based evaluation. In industrial applications of MT, task-based evaluation has most certainly been performed more frequently, but the results are typically not published.

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 to retrieving patent documents across languages (Fuji et. al, 2008). They also analysed the relationship between the accuracy of MT and its effects on the retrieval accuracy (Fuji et. al, 2009), which comes closest to the evaluation of answer retrieval in this Deliverable.

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

In the early years of NLP, already, the Verbmobil project performed end-to-end Machine-Translation as part of a longer pipeline with several modules (Jekat et. al, 2000) whereas evaluation of MT via speech-to-speech translation has been in the frame of a yearly shared task (e.g., Cettolo 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 the 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 by Schneider et al. (2010).

2. Background: HF call center application and business requirements

2.1. Monolingual workflow

The main workflow of the current (monolingual Portuguese) HF helpdesk application is depicted in Figure 1. An end user posts a question to the QA system, where it is matched against previous human generated question-answer pairs. Depending on the matching result, the next steps are:

1. If Result >= 99%, an automatic answer is sent to the client without human intervention. From HF's business perspective, this is the preferred case and long-term goal as it saves most time and money.
2. If Result >= 90%, the top 5 results are shown to an operator, who can choose to adopt one of the answers (with our without changes), or accept none of them. For HF's, this is less preferred as it saves only some time/money.
3. If Result < 90%, the operator will be asked to provide the answer with no help from the system. For HF, this is the worst option.

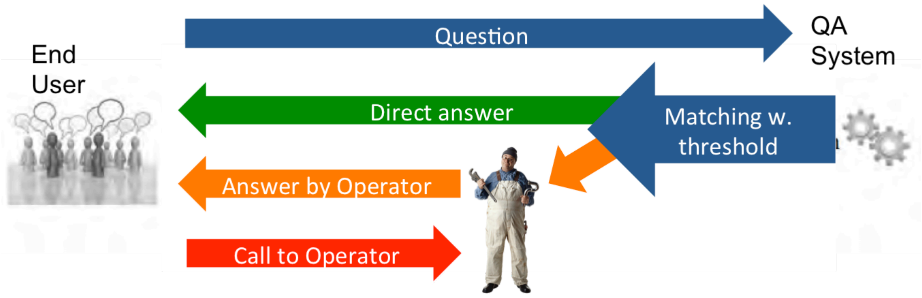


Figure 1: Main workflow of current (monolingual) usage scenario

Ideally, the answer satisfies the end user's needs and helps to solve the problem or to provide the requested information. Eventually, the user and operator will further interact via chat. In the worst case, the end user calls the operator, which consumes considerable time/money for HF that is spent on one single case.

Summing up, the two main business requirements for the monolingual scenario are:

a. Little (ideally no) human intervention in selecting the right answer (green vs. orange arrow in the figure).
b. Few (ideally no) interactions to human operators (red arrow in the figure).

2.2. Multilingual workflow

Moving to a multilingual scenario within QTLeap requires that the Portuguese database is translated to English. This process has started. 2000 Interactions have already been translated (the first two development and test sets used in setting up the QTLeap MT baselines). IDs are shared between the original Portuguese and the English database so that it is possible to compare retrieval results.

The multilingual workflow for various query languages X and the English database requires the implementation of three new main components:

1. Translation of the question from the original language to English (MT Pilots)
2. DB answer retrieval based on a "close-enough" question heuristic (HF's matching algorithm)
3. Translation of the answer from English back to the original language (MT Pilots)[1]

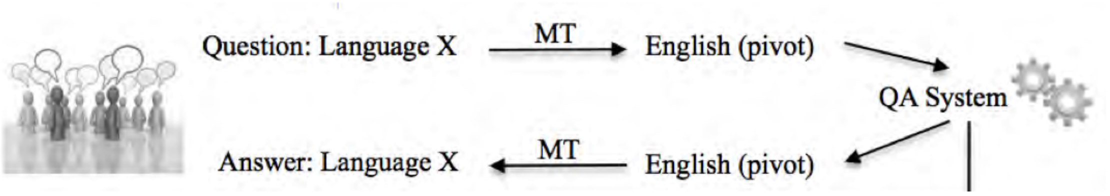


Figure 2: Simplified workflow of multilingual usage scenario

Figure 2

Figure 2 provides an overview of the multilingual scenario. As this extrinsic evaluation takes the user perspective, it is important to make clear from the beginning, who is meant by "user". Figure 3 indicates the two users within this scenario: real *end users* and the company HF as *enterprise user* of MT technology. The figure also indicates the two main checkpoints that will play a role in the extrinsic evaluation, namely (answer) *retrieval* and *publication* (translation) of the answer that has been found.

As the translation might require human intervention (answer selection, post-editing as indicated by the two humans in Figure 3), this multilingual scenario leads to additional business requirements:

[1] For efficiency reasons, the database of (English) answers should regularly be translated offline in a batch. If the QA system is not able to answer a question, a human operator will, in

c. Little (ideally no) human intervention for finding answers that are already in the database (cross-lingual answer retrieval)[2]
d. Little (ideally no) human intervention on the translated answers (post-editing)

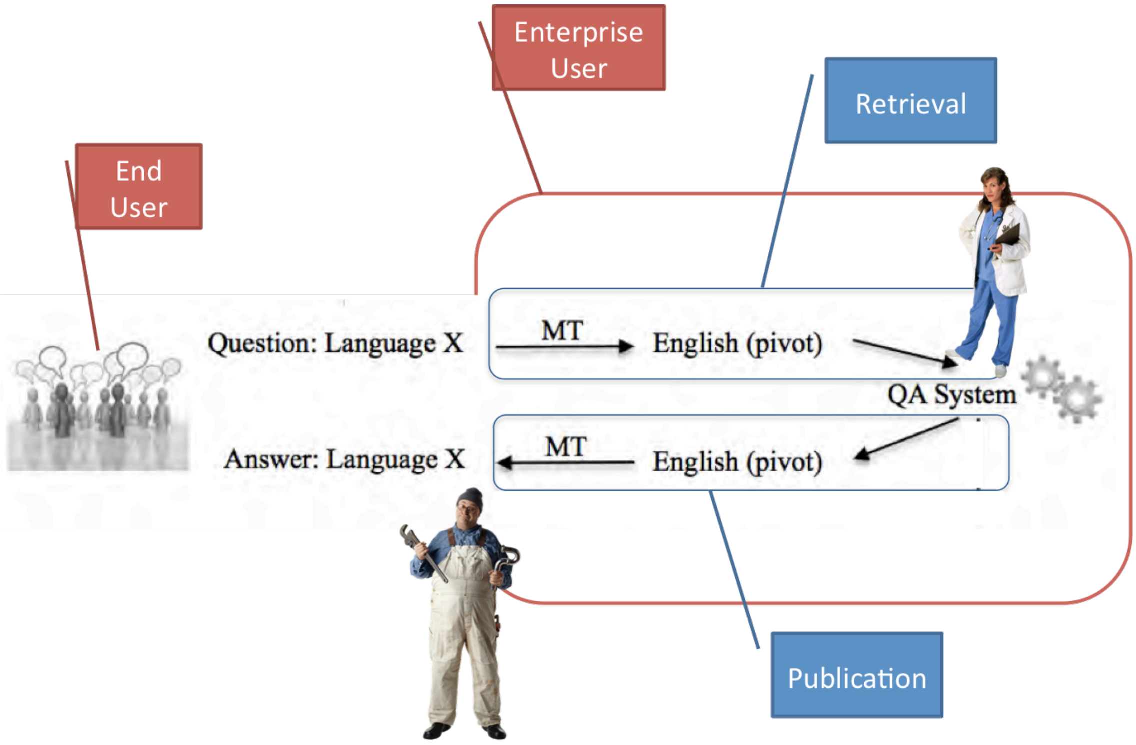


Figure 3: Simplified multilingual workflow with checkpoints indicated

3. Motivation of evaluation questions

As the multilingual service is new to HF, there is no baseline to compare its "global" performance against. The baseline would be human translation of both, questions and answers, which would not be feasible as business model.

At the same time, it would not provide useful insights to evaluate the performance of the multilingual setup (based on the smaller English database) against the performance of the monolingual setup by measuring the proportion of involvement of operators.

This would rather compare the overall performance of the two databases than assessing the specific contribution of MT. Moreover, doing a direct field test would lead to the problem that the answers would differ between settings, which would lead to too many open variables.

[2] As the project has its focus on outbound MT (publication), we ignore the case of human answer generation on the basis of a machine translated request for the time being.

Seen the above, it has been decided to perform the extrinsic evaluation in a controlled "laboratory" situation where the different factors can better be controlled. After the first evaluation round, the QTLeap team plans to revise the methodology if necessary, especially with hindsight of its relevance to "real life".

For short, what is to be evaluated is the added value of the translation, in particular if the quality of the translation is good enough to a) find and b) deliver a clear and understandable answer to final clients. The following evaluation questions have been formulated to drive the extrinsic evaluation:

i. **Retrieval**
    a. Enterprise user:
        i. **Does MT make it possible to retrieve the right answer?**
        ii. How many operators are needed?
        iii. Do they need to be bi-lingual?

ii. **Publication**
    a. Enterprise user:
        i. **How many calls to the operator need to be answered?**
        ii. How much post-editing is needed?
        iii. Can "normal" operators (that are no translators) do the editing job?
    b. End user:
        i. **Are the answers (with no post-editing) understandable & correct?**

The questions in bold face have been selected as primary questions to be answered. The next section will detail the experimental setup that has been designed for this purpose.

4. Experimental setup for extrinsic evaluation

A fundamental decision is to detach retrieval and publication in the evaluation. This will not only make it possible to identify the source of problems better, but it might turn out in the course of the project that differently optimized MT engines are to be used for retrieval and publication. Their impact on the workflow will be evaluated separately as detailed in the following sections.

The basic idea for evaluation of the retrieval module is to compare the results of the cross-lingual answer retrieval with the "gold-standard" English answers in a fully automatic fashion.

For publication, the idea is to perform a user study where the MT answer is first rated in isolation and then in comparison to a reference answer. Based on the different answers, a metric will be used to very roughly estimate the probability of calling the operator. A user satisfaction questionnaire is to be filled in at the end of the study.

The question-answer pairs for the extrinsic evaluation will (for the time being) be drawn from the translated interactions that are being produced by the project. There will be 4000 interactions used in the following way for the purpose of intrinsic evaluation:

- first batch of 1000 - evaluates Pilot 3; trains Pilot 0
- second 1000 - evaluates Pilot 0; trains Pilot 1
- third 1000 - evaluates Pilot 1; trains Pilot 2
- fourth 1000 - evaluates Pilot 2; trains Pilot 3

To avoid bias, the pool of interactions for extrinsic evaluation of the MT pilot can be formed by mixing interactions taken out of each of the four batches upon availability. This would simulate a situation where some material is "known" to the Pilot while other material is "fresh", which is plausible in a production setting like the one at hand.

For the first in vitro experiments, we intend to recruit volunteers from project partners. The goal is to evaluate 150-250 interactions per evaluation round. Later, we might decide to extend the experiments to real customers of HF or supporters from other companies of the QTLeap board of potential users.

4.1. Answer Retrieval

Starting point is the multilingual scenario where the database is in English and identical (translated) interactions in different languages share a common ID.

The goal is to determine if the use of machine translation deteriorates the results of the QA heuristic.

A list of 500 (MT translated) questions (from batch 1 and 2), the testing questions, will be evaluated in this phase. This list will be given to the QA module in order to find the correct answer for each question.

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

As has been said before, in the first evaluation, with the first MT Pilot delivered in the project, the only sensible baseline would be no translation service at all, that means that for languages not covered by the database the only solution is to contract a translator or to start to build the QA memory from the beginning. Both the alternatives are quite expensive and thus out of question. In subsequent evaluations, involving subsequengt MT

Pilots delivered in the project, previous evaluation results will serve as baseline.

In the absence of a baseline, the performance of cross-lingual answer retrieval will first be evaluated against the English reference answer(s) that can be identified via their IDs. If the translation is appropriate for this kind of task, the QA heuristic will retrieve the same answer for the same question as for the English experiment. As the QA heuristic delivers for each question a list of candidate answers along with a confidence score, it is possible to compare the results obtained with the original English questions with the one translated. It is possible to measure how far the results of translated question are from the original English question, e.g., if the retrieved answer is among the Top3/5/7/… of the monolingual heuristic, this would still require human intervention, but not writing the answer from scratch.

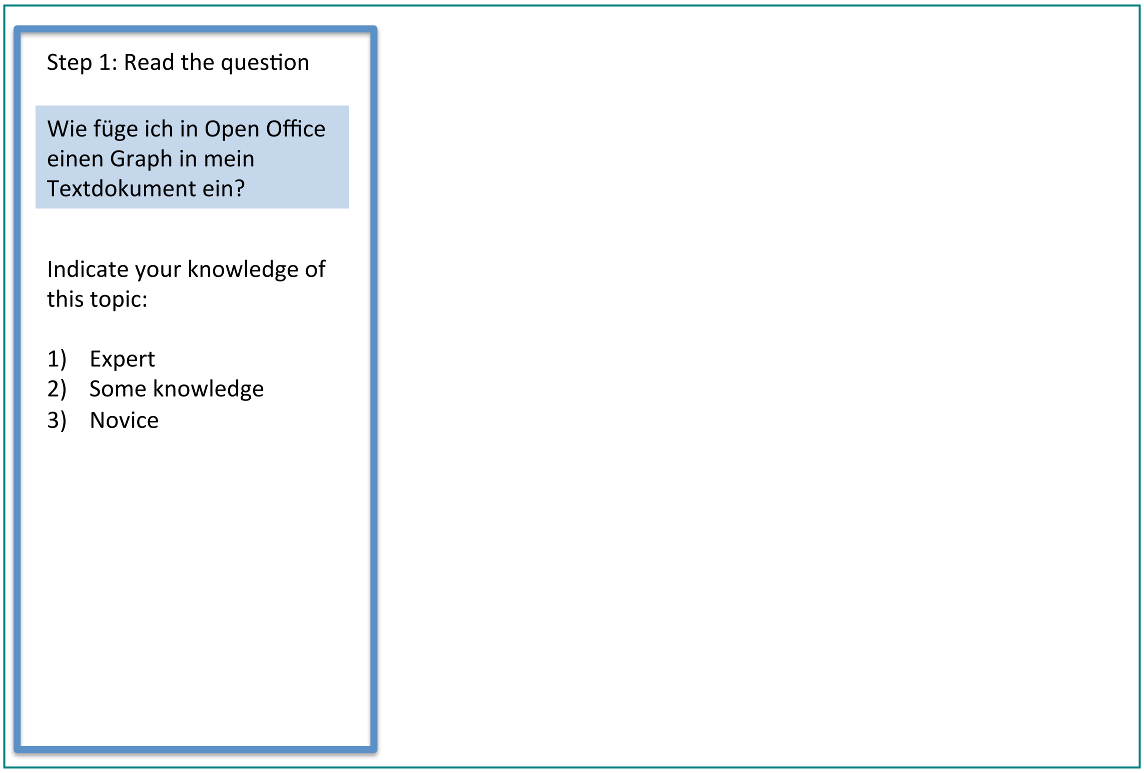
In this way we can assess the quality of translation for the specific task in a very specific context. It is possible to follow the improvement of the translation applying this evaluation for testing the different pilots.

For this particular task it is possible that a big improvement in translation quality measured with automatic metrics could not lead to a big improvement in the retrieval or just the contrary. The hypothesis is that, this type of evaluation tests aspects such as the lexical quality more than aspects such as fluency.

This kind of evaluation can be done in an automatic fashion for all the language addressed by the project for all the pilots. The only cost that could occur would be the translation of the testing questions to all the languages covered in the project if one would decide that some "fresh" questions were needed in the course of the evaluation.

4.2. Answer Publication

For evaluation of the publication stage, we intend to use a web-based interface that implements a straightforward experiment as sketched in a mockup in Image 1 to Image 3 that implements the basic idea of exposing first the MT answer and then the reference answer and have the subject evaluate the MT answer first on its own and then with respect to the reference. In order to provide insights to HF about the usefulness of MT for different customer groups, a self-estimation of the knowledge of the subject is recorded. The subjects will be instructed and in particular be warned that some of the texts they will read have been machine generated and may not be as fluent (and error-free) as a human-generated text.



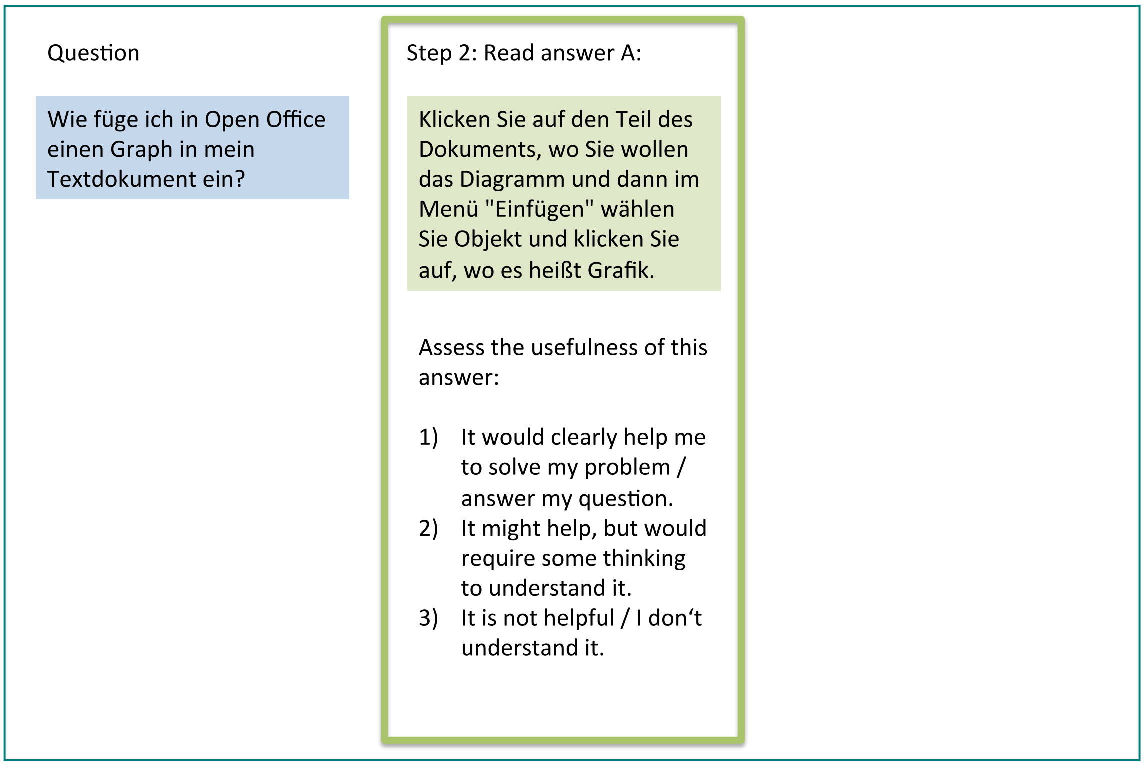
Step 1: Read the question

Wie füge ich in Open Office einen Graph in mein Textdokument ein?

Indicate your knowledge of this topic:

1) Expert
2) Some knowledge
3) Novice

Image 1: First step of mockup: exposure of question



Question

Wie füge ich in Open Office einen Graph in mein Textdokument ein?

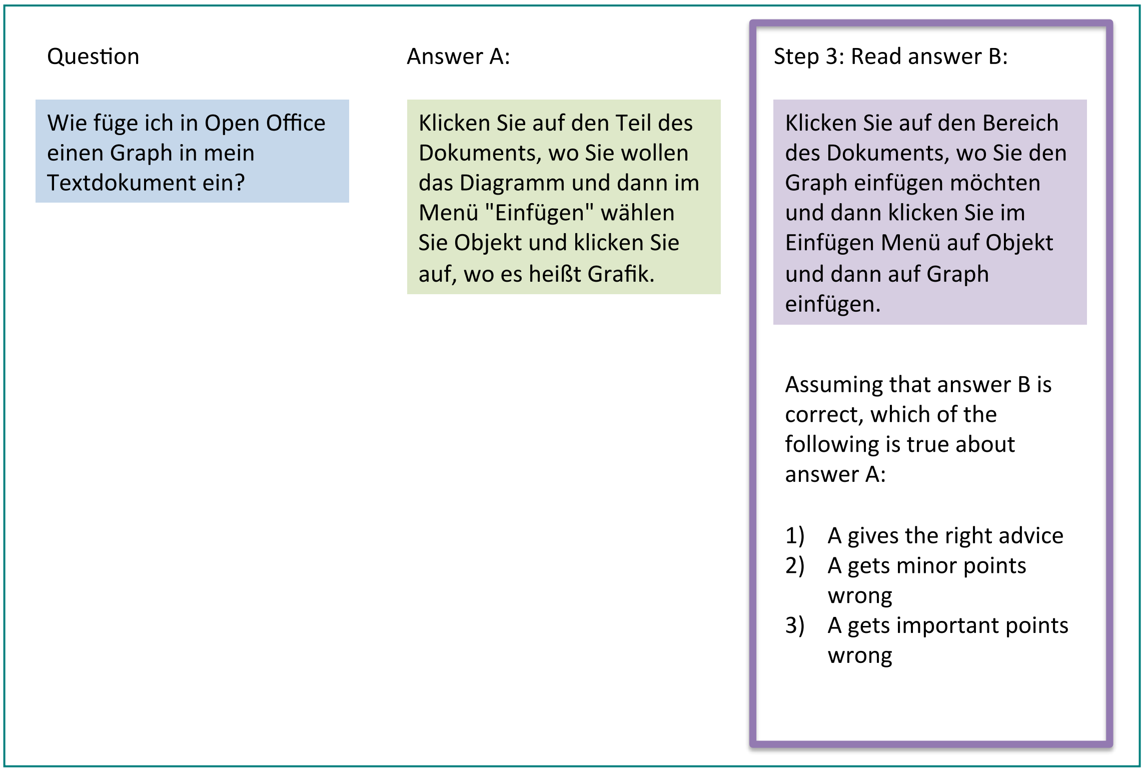
Step 2: Read answer A:

Klicken Sie auf den Teil des Dokuments, wo Sie wollen das Diagramm und dann im Menü "Einfügen" wählen Sie Objekt und klicken Sie auf, wo es heißt Grafik.

Assess the usefulness of this answer:

1) It would clearly help me to solve my problem / answer my question.
2) It might help, but would require some thinking to understand it.
3) It is not helpful / I don‘t understand it.

Image 2: Second step of mockup: exposure of MT answer and rating of usefulness



Question

Wie füge ich in Open Office einen Graph in mein Textdokument ein?

Answer A:

Klicken Sie auf den Teil des Dokuments, wo Sie wollen das Diagramm und dann im Menü "Einfügen" wählen Sie Objekt und klicken Sie auf, wo es heißt Grafik.

Step 3: Read answer B:

Klicken Sie auf den Bereich des Dokuments, wo Sie den Graph einfügen möchten und dann klicken Sie im Einfügen Menü auf Objekt und dann auf Graph einfügen.

Assuming that answer B is correct, which of the following is true about answer A:

1) A gives the right advice
2) A gets minor points wrong
3) A gets important points wrong

Image 3: Third step of mockup: exposure of reference answer and (re-)evaluation of MT answer

Based on the answers, we will then build a metric as sketched in Table 1 that allows to estimate how often a subject would call an operator. In the first row (Expert, helpful answer, right advice), the probability is low while in the last row (Novice, not helpful answer), the probability is very high. Using this matrix, different Pilots can be compared w.r.t. to the central evaluation question for publication. Moreover, it can serve as input to intrinsic evaluation, e.g., it can be interesting to inspect those answers that are marked as difficult to understand in step 2.

Step 1	Step 2	Step 3	Probability of calling operator
1)	1)	1)	low
1)	2)	1)	low
2)	2)	2)	medium
	.		
	.		
	.		
3)	3)	--	high

Table 1: Metric for estimating operator call probability (incomplete)

In subsequent evaluations, it is planned to either phase in real end users as subjects or to "simulate" real user interest by asking subjects to pre-select questions that are meaningful to them in the sense that they can decide if the answer they got would have helped them. A questionnaire assessing the user satisfaction with several targeted questions based on a Likert scale will then be handed out. A first version of this questionnaire based on Schneider et. al (2010) is shown in the appendix. It will be refined according to HF's needs in the course of the actual evaluation.

5. Time line

The following Deliverables provide an overview of the upcoming integration and evaluation tasks:

M11: Embedding of a baseline MT into the QA system (D3.5)

M12: Report on embedding and evaluation of a baseline MT (D3.6)

Revision of extrinsic evaluation methodology

M17: Embedding of the first MT pilot into the QA system (D3.7)

M18: Report on embedding and evaluation of the first MT pilot (D3.8)

M24: Embedding of the second MT pilot into the QA system (D3.9)

M24: Report on embedding and evaluation of the second MT pilot (D3.10)

M34: Embedding of the third MT pilot into the QA system (D3.11)

M35: Report on embedding and evaluation of the third MT pilot (D3.12)

6. References

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th iwslt evaluation campaign. In Proceedings of the 10th International Workshop on Spoken Language Translation, pages 29–38, Heidelberg, Germany.

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, pages 429–433.

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 Proceedings of the 7th TCIR Workshop Meeting, pages 389–400.

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 Press.

Susanne J Jekat and Walther v Hahn. 2000. Multilingual Verbmobil-dialogs: Experiments, data collection and data analysis. In Verbmobil: Foundations of Speech-to-Speech Translation, pages 575–582. Springer.

Anne Schneider, Ielka Van Der Sluis, and Saturnino Luz. 2010. Comparing intrinsic and extrinsic evaluation of MT output in a dialogue system. In Proceedings of the 7th International Workshop on Spoken Language Translation, pages 329–336, Paris.

H.-C.a Wang, S.R.b c Fussell, and D.b Cosley. 2013. Machine translation vs. Common language: Effects on idea exchange in cross-lingual groups. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pages 935–944.

Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In Proceedings of ACM CSCW'06 Conference on Computer-Supported Cooperative Work, pages 515–524.

7. Appendix: User satisfaction questionnaire

		Strongly disagree			**Neither**		**Strongly agree**	
	Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	I could quickly find what I was looking for.							
2	I had serious problems understanding the texts.							
3	There were awkward words and phrases in the dialogue.							
4	The utterances were fluent.							
5	I would rate the utterances as incomprehensible.							
6	I would rather use the English original.							
7	I would prefer to getting help calling an operator							
8	The system responses agreed with my expectation.							
9	The system did always understand what I said.							
10	The system did not give me enough information.							
11	The system gave me a lot of unnecessary information.							
12	The system's responses were appropriate.							
13	The system gave me too much information in one go.	7.1.1						
14	I would consider using a similar system to ask for help in a similar context.							
15	I could quickly find what I was looking for.							

Table 2: Questionnaire based on a 7-point Likert scale