Artificial Cognitive MT Post-Editing Intelligence

Jörg Schütz

Bioloom Group and Saarland University

Bahnhofstraße 12

D-66424 Homburg

Germany

`joerg@bioloom.de and j.schuetz@mx.uni-saarland.de`

Abstract

Post-editing (PE) is a necessary process in every MT deployment environment. The competences needed for PE are traditionally seen as a subset of a human translator's competence. Meanwhile, some companies are accepting that the PE process involves self-standing linguistic tasks, which need their own training efforts and appropriate software tool support. To date, we still lack recorded qualitatively and quantitatively PE user-activity data that adequately describe the tasks and in particular the human cognitive processes accomplished. This data is needed to effectively model, design and implement supportive software systems which, on the one hand, efficiently guide the human post-editor and enhance her cognitive capabilities, and on the other hand, have a certain influence on the translation performance and competence of the employed MT system. In this paper we argue for a framework of practices to describe the PE process by correlating data obtained in laboratory experiments and augmented by additional data from different resources such as interviews and mathematical prediction models with the tasks fulfilled, and to model the identified process in a multi-facetted fashion as a basis for the implementation of a human PE-aware interactive software system.

1   Introduction

To deal with the increasing complexity of the human society within a global economy, artificial cognitive systems will play an ever increasing important role in the near future. Today, although a lot of new insights in human cognition and the processes within the neocordial system have advanced, our understanding on how the human brain operates is still in an early stage. Some recent studies assume that the human brain is a huge memory system which stores or even models experiences in a way that reflects the structure of the world by remembering sequences of events and their relationships, and making predictions based on those memories initiated or triggered by different external stimuli. It is believed that this memory prediction system forms the basis of intelligence, perception, creativity, and even consciousness (see e.g. Hawkins, 2004).

Post-editing of MT output is an intrinsic part of every MT deployment scenario to guarantee a publishable quality and overall consistency of the eventual translation result – except for gisting in which MT raw output is used for quick insides into the foreign language message – and is independent of the MT type employed in the automatic translation.

There have been various in-depth investigations to define the process of post-editing at various levels, see for example Krings (2001) and O'Brien (2006a and 2006b), and in industry such process-oriented research is mainly coupled with the aim to identify an economic basis to calculate and to reduce the cost of PE.

Nevertheless, the actual cognitive workload and the mental conditions of the human post-editor have never been the subject of any detailed research. This situation is changing now, and recently an interdisciplinary group of European industrial and academic researchers has proposed an R&D

project in this direction within the Seventh Framework Programme of the European Commission.

In this paper we argue for the introduction of a general framework of practices which firstly allows to gain sufficiently detailed data from monitoring the human PE process, secondly to derive process models from analyzing and interpreting the data, and thirdly to provide a theoretically sound basis for designing and building software systems that support the human post-editor in her tasks and activities.

In addition, the framework should also enable us to establish and implement methods on how an employed MT system might benefit from an artificially guided PE process with proactive feedback cycles by implanting emergent behavior into these process models.

The proposed framework of practices is based on a combination and an intertwining of different interdisciplinary scientific approaches with the aim to open up new ways of combining machine learning, knowledge-driven and cognitive techniques and methodologies to language understanding, generation and translation by humans and machines. We expect to contribute and also to improve our understanding of the mechanisms underlying artificial and human cognition, and thus enabling us to build systems that can better anticipate, learn and emulate effective human behavior, and reflect it in efficient human-machine interaction by providing a robust and versatile collaborative platform.

The research described here is work-in-progress, and is therefore meant to stimulate the discussion within the MT community and to foster broader community research efforts.

2 The PE Process

The PE process mainly consists of different repair activities within a 2-source scenario – source text and raw MT output or translation memory data – as opposed to the 1-source scenario of human translation.

In the 1-source scenario the translator's focus is entirely on the source text and the production of an adequate and fluent target text taking into account certain constraints on language and cultural levels.

In the 2-source scenario the post-editor has to switch between the comprehension of the source text and the MT output, which for the MT output might involve inferencing or entailment activities to effectively interpret the raw target text and to identify the appropriate repair action.

For this reason, most companies maintain internally explicit PE guidelines to minimize the PE resources and cost but still guaranteeing a sufficient translation quality in a minimum of time.

The most obvious tasks in the 2-source scenario on the MT/TM output are:

- replacement of unknown words
- deletion of superfluous translation alternatives
- correction of lexical and grammatical errors with a particular focus on terminology
- revision and rewriting of parts of or entire sentences

These tasks are very much in-line with the topics of the translation quality measures that have been defined jointly by several automotive manufactures as the SAE J2450 recommendation and standard. The rational behind J2450 is that the automotive industry is one of the industries which has to maintain a large number of translation languages – in some cases up to 32+ languages. Other industries, e.g. software, pharmaceutical and life-sciences industry, have introduced similar measures.

Because the cost-efficiency of any PE service is the most valuable driving force in industry, the set-up of any PE process must fulfill the following basic conditions:

- pre-assess the MT suitability
- limit the PE time to approximately 15 to 20 minutes per ordinary print page
- keep PE as a part-time task for translators
- ensure a positive attitude of post-editors towards MT technology, and train them in understanding how MT works

The last condition was only recently taken into account, and mostly with the introduction of Authoring for Translatability technologies such as CLAT of IAI (IAI, 2008) and the IQ Suite of acrolinx, (acrolinx, 2008) which also need thorough training phases, and which are mainly employed in the industries mentioned above.

The PE process normally takes place in specific PE software environments, and is integrated into the overall business process landscape to allow for effective monitoring and time optimization operations. In these PE environments the post-editor either works on a paragraph level according to the

structural markup specifications of a document, or on a sentence level depending on the MT segmentation capabilities. In each case, the post-editor has the option to view a broader preceding or subsequent context of the actual focussed segment.

For our framework of practices we make the following assumptions:

1. the PE process and its tasks are different from the translation process and its tasks
2. PE is becoming very important in different industrial supply chains and associated operations as well as in self-service support
3. so far no coherent PE user-actvity data (UAD) is available along different dimensions
4. PE involves functional text production, interlingual communication skills, and an understanding of the "MT jargon"
5. "train learning" is an essential paradigm for future human and machine interactions in both directions

3   PE Process Data Acquisition

3.1   General Method

To gather UAD the think-aloud method and protocol has been preferred in several task analysis settings including PE (see Krings, 2001 and O'Brien, 2006a). However, in settings that need high mental capabilities and cognitive resources as for our PE case, the method has extreme negative effects, and might cause over 25 % delay in the actual processing.

Therefore we encourage to base the initial data acquisition step – resulting in raw data – on a method that automatically observes and logs eye movements and records keystrokes in a laboratory PE environment. The idea is that in a stochastic data analysis phase – resulting in annotated or cooked data – we can directly correlate along a time line the raw data with the identified human activities associated with the process and its different tasks:

• reading and understanding the two sources
• producing the final translation including replacements, deletions, corrections, revisions and rewritings

Before we can actually analyze the logging data, we also have to define an appropriate method for the data interpretation, and to decide on how to represent the information (marked-up data). The

markup language is based on a vocabulary maintained in an ontology to allow for effective data interchange with other groups working on similar research tracks.

Marked-up data is then further used in machine learning sessions (supervised and unsupervised) for purposes such as activity clustering and categorization of tasks and actions. To accomplish this specification we first create training data derived from interviewing post-editors in replay sessions. In a replay session interview the post-editor makes explicit how her tasks and actions have been executed within the PE process. This information augments the cooked data in terms of descriptive elements for each PE task and action. Any prognoses detected in the annotated data sets are fed back into new laboratory data acquisition cycles.

The data acquisition approach is designed open so that additional data from e.g. fMRI (functional Magnetic Resonance Imaging) analyses can augment and extend the data acquisition cycles at a later stage.

3.2   Eye Movement Tracking

Eye movement tracking is a method that has been successfully applied in general GUI design and in the evaluation and assessment of human-machine interfaces – GUI usability testing – as well as in designing barrier-free access to Web pages – web site usability testing. So far research on real-time usage of this type of data to establish and to enable assistive adaptation of human-machine interactions has been limited to certain intelligent learning environments and problem solving strategies. Prominent in these scenarios is in particular the analysis of a human's reading perception.

Although the basic facts about eye movements have been known for almost one hundred years, only recently research activities have started to look at eye movement behavior as a reflection of the human cognitive processing. Let us exemplify this with the reading task, which is one of the activities in the post-editing scenario: skilled readers, and we assume this fact for the human post-editor, move their eyes during reading on the average of every quarter of a second. During the time that the eye is fixated, new information is brought into the mental processing system by an external stimulus. The average fixation duration is approximately 200 ms to 250 ms, the actual range is from 100 ms

to over 500 ms, and the distance the eye moves in each saccade, i.e. a short rapid movement, is between 1 and 20 characters with the average being 7 to 9 characters. The most important fact about fixations, i.e. the eventual focal point of a saccade, and saccades is that there is considerable variability not only between readers, but for the same human who reads a single segment of text. In addition, skilled readers make regressions back to material already read about 15 % of the time. If the perceptual span observed includes all or many of the words in a segment of text, then eye movement measures would not reveal much information about the cognitive processing. However, if we assume that the reader gains only useful information from the word directly focused on, then eye movement behavior contributes essentially on what role the eyes play in the recognition (understanding) and translation process.

Because of these observation it is necessary to augment the logged eye movement data with the post-editors explanations in a subsequent replay session interview. The interviews are also necessary to prove our assumption on the role of the visual stimuli in the PE process.

3.3 Keystroke Logging

Keystroke logging is the method to study how the post-editor interacts and accesses with the PE system environment. In addition, the data obtained is also useful to measure the user productivity on the different tasks involved in the PE process which reflects to certain extent the amount of the cognitive processing as well as the mental workload. In this sense, keystroke logging brings in yet another quality of raw and cooked data thereby influencing the prognosis phase.

3.4 Laboratory Environment

The laboratory environment presents the technical equipment for the PE test sessions consisting of video and eye tracking equipment, and the software for the PE, MT and keystroke logging systems.

Most important in the laboratory environment is the careful selection of a homogenous group of human subjects – post-editors – who must share similar linguistic, educational and professional backgrounds to allow us to assume a certain homogeneity in the subjects' linguistic competence and per-

formance, and thus their potential responses to the MT output with which they are faced during the tests.

The setup of the laboratory environment is inspired by environments used in medical evidence-based approaches to pre-clinical medication testing.

4   Data Analysis

The ultimate goal of the PE user-data analysis is to detect correlations between the recorded logging data (in the fMRI case: the detected brain activations) and the task the subject reports to perform during the observation (the fMRI scan).

4.1   Annotated PE Data

Because we are dealing with huge amounts of data, we have to identify the necessary data elements that allow for a sufficiently detailed process description at the task and activity level. This means that we have to break or splice the data streams into useful clusters which are categorized by tasks and single actions. Machine learning techniques are employed in this step, and the results are validated against new raw data from continuous laboratory tests. This cycling is necessary because the monitored data is not static, and variations in the subject's behavior have to be taken into account. The same is true for the MT output, and therefore we also have to take care on variation resulting from the automated translation.

The identification of the different parameters that form the set of discriminating properties in the data streams are used as the essential markup elements for the annotated PE data.

Additional properties that augment the annotated PE data comes from the interviews, which however need to be integrated manually. This information includes task and action assignments as well as data derived from observed result variations of the eventual translation output. We might represent these variations in Choice Networks as this was suggested in previous works by Krings (2001) and O'Brien (2006b). The benefit of the CN representation is that they accomplish for language related aspects in the PE process, and which might be useful in extracting data for the MT feedback operations.

The raw data based approach has the advantage that it is less resource intensive because we do not

integrate explicit human knowledge. However, in general the raw data based generated prediction model is a black box which does not allow to understand the relationship between the employed data and the model. Therefore, we also want to investigate how a reduction in model information affects qualitatively the assistive behavior of the envisaged artificial cognitive PE environment's interaction adaptation.

The initial training data needs a large amount of manually polishing and revision cycles but in subsequent training and learning phases it is an automated process. These phases deliver the prediction model(s) of our framework of practices, whereas the PE process as such is modelled in a standard Business Process Modeling (BPM) suite.

4.2 Process Modeling

The PE process modeling phase is accomplished similar to the modeling of business processes because we envisage to employ standard BPM tool suites. This approach enables us to subsequently implement the PE tasks and actions as different software services. The semantics of the process data is also maintained in the framework's ontology.

These software services form the executional level of our eventual artificial cognitive MT PE intelligence.

5 Artificial PE Intelligence

5.1 Cognitive Architecture

In our framework a cognitive architecture is the blueprint for intelligent PE agents. It proposes artificial computational processes that in some sense act like human cognitive systems, i.e. our human post-editor, or act intelligent under some definition – task or action oriented. This cognitive architecture forms a subset of the general software service architecture. The architecture models not only behavior, but also structural properties of the modelled system. Because these properties need not be physical properties, they can be properties of virtual machines implemented in physical machines e.g. in brains or computers.

Furthermore, within the framework we define and qualitatively evaluate how different resources should be developed in order to better match with the cognitive processes of the post-editor.

Because the monitored data resulting from our experiments is clustered and categorized along different dimensions such as associated task and action, we expect to additionally verify how important it is to have other fine-grained information – evidence-based validation – on the post-editor's high-level mental states within the various types of interaction supported by the post-editing environment.

This information is an intrinsic part of the cognitive architecture, and is based on the two prediction models that we have established within the described learning phases on the same type of system and process environment: (1) the knowledge-based prediction model – supervised technique – , and (2) the raw data prediction model based on an entire raw data based approach – unsupervised technique.

5.2 Direct PE Support

The laboratory work focusses entirely on the tasks and activities of the post-editor without interrupting her activities, and will be the basis to develop machine intelligence that allows to automatically learn from and to adapt to the behavior of the human actor.

This includes guiding the process effectively along the lines of the identified process model by highlighting the identified tasks and its associated actions, thereby avoiding repetitive tasks.

This research will also reveal important insights about the similarities and the differences between human and artificial intelligence, and about the problems encountered when human and artificial intelligence have to effectively and inefficiently interact with each other.

5.3 Indirect MT Support

In addition to the human activity data, the data analysis phase also allows us to gain insights into the MT system performance and its translation competence. The MT output is not only assessable according to errors or misbehavior in terms of e.g. wrong analyses and generation, lexical and terminological gaps, inadequate language models, etc. but also in terms of possible adaptive training methods to revise the underlying performance models of the MT agent.

Within our framework of practices this opens up the perspective to closer investigate, and even model and develop methods that would effectively support and train the MT system on how to adaptively improve its output quality, and as a consequence, for example, to avoid repeating the same mistakes.

This research track, however, needs a close interaction with the MT system developers as well as their willingness to be engaged in such a research enterprise. Preliminary investigations in this direction seem to justify that hybrid MT systems that combine statistical approaches with linguistics-based approaches are very well suited for these types of investigations.

Our view is that resource bottlenecks combined with low performance, inappropriate adaptivity and general language competence of an MT system have been traditionally a major obstacle to novel applications for interpersonal communication involving automated translation across application scenarios.

As a major result, such a collaboration should expect to design and implement a new breed of MT system together with a post-editing environment that will be able to appropriately solve several problems and shortcomings of currently existing technologies.

6 Conclusion and Perspectives

Our solution of building a framework of practices for an artificial PE intelligence with its own intrinsic metabolism is inspired from different fields including cognitive systems, neurosciences, computer sciences and information bionics. As such the envisaged PE software system also shares properties of biological systems, namely emergence, robustness and modularity.

Emergence is a trait in which the whole is greater than the sum of the parts, which in our scenario is accomplished through the employed techniques and technologies for dealing with and abstracting upon different but-related information elements – including the MT system.

Robustness is characteristic of resilience to fluctuations in the immediate environment resulting from redundancy and control mechanisms, which we have implemented with our data-driven performance control strategy on different levels.

Modularity is a phenomenon that explains the clustering of parts into a functional or structural entity, which we have achieved by modeling the process knowledge through an ontology acting as a multi-dimensional logical task and action index that equally addresses all data and information resources.

The realization of a reasonably performing new post-editing system is certainly a significant breakthrough economically, socially, scientifically as well as technologically.

We expect that our framework of practices will influence the future creation and establishment of enhanced and new language related products and services as well as the design and development of novel adaptive human-machine interfaces including possible direct interfaces between artificial artifacts and living systems which today in science is termed Bio-ICT convergence.

References

Acrolinx. 2008. Web site: http://www.acrolinx.de.

IAI. 2008. Web site: http://www.iai-sb.de.

Hans P. Krings. 2001. *Repairing Texts:* Empirical Investigations of Machine *Translation Post-Editing Processes*. Edited/translated by G.S. Koby. The Kent State University Press, Kent, Ohio, USA.

Gyde Hansen. 2003. *Controlling the Process: Theoretical and Methodological Reflections on Research into Translation Processes*".In: Fabio Alves (ed.) "Triangulating Translation". John Benjamins, Amsterdam, The Netherlands. pp. 25-42.

Jeff Hawkins and Sandra Blakeslee. 2004. *On Intelligence*. Henry Holt and Company, New York, USA.

Matthew MacDonald. 2008. *Your Brain: The Missing Manual*. O'Reilly, Sebastopol, CA, USA.

Susana Martinez-Conde, Stephen L.Macknik and David H.Hubel. 2004. The Role of Fixational Eye Movements in Visual Perception. In: Nature Reviews, Neuroscience, Vol. 5. Nature Publishing Group, New York, USA. pp. 229-240.

Sharon O'Brien. 2006a. *Controlled Language and Post-Eding*. In: Multilingual Writing for Translation Guide. Multilingual Computing Inc. pp. 17-19.

Sharon O'Brien. 2006b. *Machine-Translatability and Post-editing Effort: An Empirical Study using Translog and Choice Network Analysis*. PhD Thesis, School of Applied Language and Intercultural Studies, Dublin City University, Dublin, Ireland.

Stephan Schleim. 2008. *Gedankenlesen, Pionierarbeit der Hirnforschung*. Heise Verlag, Hannover, Germany.