[From: *Tralogy*, Paris, 3-4 March 2011]

John Moran and David Lewis

Unobtrusive methods for low-cost manual evaluation of machine translation

Abstract

Machine translation (MT) evaluation metrics based on n-gram co-occurrence statistics are financially cheap to execute and their value in comparative research is well documented. However, their value as a standalone measure of MT output quality is questionable. In contrast, manual methods of MT evaluation are financially expensive. This paper will present early research being carried out within the CNGL (Centre for Next Generation Localisation) on a low-cost means of acquiring MT evaluation data in an operationalised manner in a commercial post-edited MT (PEMT) context. An approach to MT evaluation will be presented which exposes translators to output from a set of candidate MT systems and reports back on which system requires the least post-editing. It is hoped that this approach, combined with instrumentation mechanisms for tracking the performance and behaviour of individual post-editors, will give insight into which MT system, if any, out of a set of candidate systems is most suitable for a particular large or ongoing technical translation project. For the longer term we propose that post-editing data gathered in a commercial context may be valuable to MT researchers.

Introduction

In a recent survey carried out by SDL[1] in conjunction with the Association for Machine Translation in the Americas (AMTA) and the European Association for Machine Translation (EAMT) 30% of the respondents working as managers in large corporations surveyed indicated that they are using or plan to use PEMT.[2] In particular technical documentation is considered “prime content” across the sectors surveyed. Interestingly, over 40% of respondents were more likely to use PEMT than recorded in a similar survey two years prior.

For the long term, we believe that this kind of commercial post-editing data provides intuitive and valuable insights into the quality of MT. Particularly, when open-source MT systems are used to generate baseline quality MT in commercial translation projects, post-editing and other instrumentation can provide a means to evaluate those algorithmic improvements manually. Industrial post-editing data can also be used to verify that automated metrics correlate with human PEMT effort in various contexts. However, algorithmic improvements can only be verified if other factors can be normalised. SMT system performance, the dominant MT paradigm, is highly dependent on training corpora. Efforts are currently underway by organisations such as the Translation Automation User Society (TAUS[3]) to gather large specialised bilingual corpora to train

MT systems. These corpora can be combined with large internal translation memories (TMs).

The combination of these sources of linguistic training data with commercial or open-source SMT toolkits like Moses (Koehn et al., 2007) and Joshua (Li et al., 2010) suggests that the current surge in PEMT usage for technical translation is unlikely to abate in the near future.

However, recent research has shown that translators are highly variable (20% to 131%) in terms of PEMT productivity improvements (Plitt and Masselot, 2010). Unfortunately, current methods[4] for comparing human translation (HT) with PEMT and for comparing MT systems with each other require much manual intervention. Our system is designed to minimise intervention on the part of translation project managers (or localisation engineers) when measuring PEMT productivity in large multiple-translator translation projects. Although our system is designed to fulfil a commercial function (measurement of PEMT translator productivity), we hope that in the future our system could also be used to gather manual MT evaluation data on a larger scale than is currently possible.

Overview

We begin our discussion by situating our research approach within the context of other approaches to post-editing research by describing how granularity of data analysis is traded off against volume of sample sets of sentences.[5] We describe how we intend to address the granularity of analysis issue by means of instrumentation of computer-aided translation (CAT) tools to track individual post-editing behaviour and speed. We also describe how we intend to address the volume of analysis issue by proposing the use of a translation management system (TMS) to keep track of translation project variables. We then discuss a growing sentiment that improvements in MT are becoming harder to measure using current automated techniques. We content that innovations in techniques to gather manual MT evaluation data may go some way to solving this problem.

We begin our summary of the current state of the art in elicitation of manual MT evaluation data by listing four current approaches. We briefly discuss two previous PEMT studies and discuss a third in some detail (Plitt and Masselot, 2010), where a custom-made web application was used to record post-editor speed to evaluate PEMT productivity compared with human translation (HT) productivity while blind-testing for quality using an open-source statistical MT (SMT) toolkit. We outline how our approach differs from theirs: translators using our system will use a CAT tool, we will use a translation management system (TMS) to record project variables and we seek to compare MT with MT as well as PEMT with human translation (HT). We then provide a more thorough explanation of why we were motivated to adopt our approach and why we chose to use three existing software systems, namely OmegaT6 and Trados Translators Workbench™[7] (Trados) as CAT tools and GlobalSight as a TMS. Finally, we outline a design for a software architecture which connects these components to an MT brokering system so that HT can be compared to PEMT, and MT systems can be compared to each other (as pairs or with many candidates). We conclude by discussing next steps we intend to carry out in our research.

Post-editing research

Peer-reviewed articles on the topic of MT post-editing are not as common as articles published on the subject of MT itself (S. O'Brien, 2005). However, there is enough research on the topic to ascertain some general trends in the approaches taken. One of these patterns is a classic trade-off scenario between granularity of analysis and volume of analysis, as illustrated in Fig. 1 below. The x-axis denotes granularity of analysis of the post-editing task, and the y-axis denotes the volume of sentences analysed. For example, analysis of post-editing using eye-tracking equipment and a specialised post-editing logger necessarily involves fewer sample sentences than a study which looks at post-editing from a human resource perspective, where daily translation output rates are measured.

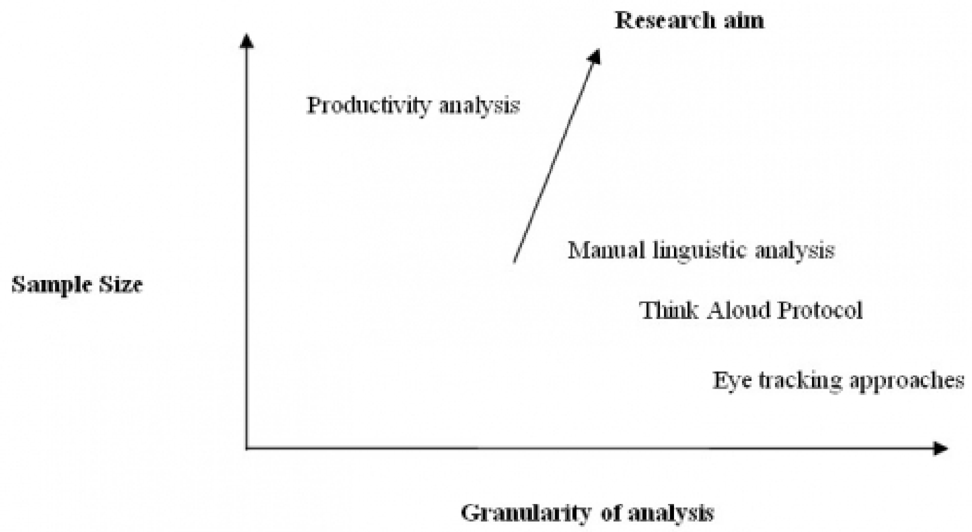

Fig. 1

In particular, translator productivity analyses carried out in the context of commercial projects generally analyse a greater number of sample sentences, because they can elicit their evaluation data from linguistic experts (translators) who are being paid for the task outside of the context of the study.

Some studies which fall into each category in Fig. 1 in order of typical sample data set size are:

- Translator productivity analysis, for example, (Allen, 2004), (Groves and Schmidtke, 2009), (Plitt and Masselot, 2010), (Guerberof, 2009)
- Manual linguistic analysis, for example, (Tatsumi, 2009)
- Think aloud protocol (TAP) approaches, for example, (Krings, 2001)

- Eye-tracking approaches, for example, (Doherty and O'Brien, 2009)

Although more highly granular manual linguistic analysis and eye-tracking approaches have yielded interesting and promising results, this paper is mainly concerned with translator productivity analysis in the context of PEMT. The aim of the research presented in this paper is to provide a somewhat greater granularity of analysis for a much greater number of post-edited sentences without causing an explosion of effort. This is illustrated by the diagonal arrow labelled "research aim" pointing up and (slightly) to the right.

The problem with automated evaluation metrics

Automated MT metrics like BLEU, METEOR or TER[8] are an indirect measure of MT quality, as they are based on an observed correlation between manual MT evaluation and various string distance algorithms which measure the difference between a[9] reference manually translated sentence and a machine translated version of the same source sentence. These comparisons are normally carried out over a corpus of test sentences.

BLEU[10] is an early example of such an automated evaluation metric which was shown to have a high correlation with human judgement. This was later confirmed by (Coughlin, 2003). However, when presented by (Papineni et al., 2002) the metric was never intended as a replacement for human evaluation. Indeed, this point was highlighted in their abstract,

"We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations."

More recently, this position is stated in (Callison-Burch et al., 2010).

"It is our contention that automatic measures are an imperfect substitute for human assessment of translation quality."

If we make a distinction between MT research in general and the application of automated evaluation metrics to a specific PEMT project, we can see that the problem becomes more acute.

"In our experience raw-MT evaluation metrics such as BLEU (Papineni et al., 2002), and human evaluations and ratings of smaller samples, are not in themselves sufficient or representative indicators of how useful MT will be on specific larger post-editing projects."(Groves and Schmidtke, 2009)

There can be no doubt that advances in automated MT evaluation metrics have, over the past decade or so, contributed to significant incremental improvements in the field of MT, and we do not dispute their value during MT system development. However, the high financial cost of carrying out manual evaluations combined, we suspect, with the success of automated methods may have led to an overreliance on automated evaluation metrics when presenting and evaluating novel algorithmic techniques designed to improve MT output quality.

This is commented on in passing by (Coughlin, 2003), who states that,

"...practical considerations have forced the field to rely on automated metrics..."

As large comparative improvements using automated metrics become increasingly hard to achieve (in particular for well-studied language pairs), automated metrics have come under some criticism in recent years.

The quality of the MT output does not correlate any more with the MT evaluation metrics that we use today. About five years ago the MT evaluation metrics that we had were better than the MT systems that they were evaluating but I think the MT systems now have outgrown those string based evaluation metrics...[11](transcribed interview with Way, 2009)

MT metrology research seeks to address this issue by focusing on improving the correlation between human evaluation scores by various means.

"In order to be both effective and useful, an automatic metric for MT evaluation has to satisfy several basic criteria. The primary and most intuitive requirement is that the metric have very high correlation with quantified human notions of MT quality." (Banerjee and Lavie, 2005)

The results of tests to measure the correlation between various automated metrics and human assessments of MT quality as part of the MetricsMATR competition are described in (Callison-Burch et al., 2010).

As we can see, MT metrology research itself relies on manual evaluation scores, so we keep returning to the same problem. It is expensive to gather MT manual evaluation data on a large scale. In the remainder of this paper we describe current methods to gather manual MT evaluation data and suggest an innovative approach to gathering data from commercial PEMT projects as a means towards solving this problem.

Four current approaches for gathering MT manual evaluation data

In this section we will examine four methods which summarise the state of the art for manual MT evaluation data collection.

- Shared tasks competitions
- Asking volunteers
- Paying experts or non-experts in a classic, non-commercial experimental context or via crowdsourcing[12]

- Unobtrusive and obtrusive methods for measuring post-editing behaviour in commercial translation projects

Shared task competitions

Shared task competitions are common to a number of fields in computer science. They provide a means for researchers to prove the value of their work relative to comparable systems.

In the field of MT competitions like WMT[13] (Workshop on Statistical Machine Translation) and IWSLT[14] (International Workshop of Spoken Language Translation) make a valuable, and we believe irreplaceable, contribution to MT evaluation by collating evaluation results, levelling the playing field and taking advantage of natural ergonomic synergies to reduce the cost of evaluation.

As we are mainly concerned with written language translation, we will focus our attention of the WMT shared task evaluation competition, which has been held annually for the past five years. Its primary objectives are to

“...evaluate the state of the art in machine translation, to disseminate common test sets and public training data with published performance numbers, and to refine evaluation methodologies for machine translation.” (Callison-Burch et al., 2009)

Some of the costs involved in running the competition are described by the organisers.

“The total cost of creating the test sets consisting of roughly 80,000 words across 3027 sentences in seven languages was approximately 31,700 euros (around 39,800 dollars at current exchange rates, or slightly more than $0.08/word).” (Callison-Burch et al., 2009)

This was the cost of employing professional translators via a translation agency, so it does not take into account paid annotators. However, the cost was deemed acceptable, as 87 MT systems and 22 system combination entries were evaluated. At least, on this shared task basis the cost was easier to justify than for a single MT system. We believe that shared task competitions make a valuable (we believe, irreplaceable) contribution to MT evaluation by collating evaluation results to measure the state of the art on a level playing field and taking advantage of natural ergonomic synergies to reduce the cost of MT evaluation. However, competitions like the WMT have a finite budget, so they too have some inherent restrictions:

- In its current form each WMT competition focuses on only one genre.
- In the case of WMT 2009, cited above, this was a written news 15corpus.
- They are held infrequently (for example, the WMT is held on an annual basis)[16].
- They can evaluate only a finite number of language pairs.

- Labelling is carried out by WMT contestants, who are expected to carry out 8 hours of human evaluation. However, for PEMT, professional translators are the target user group, so care must be taken with the word “expert”.

Asking volunteers

A common approach used to gather manual evaluation data is to ask volunteers to provide MT evaluation data. In general, it is difficult to gather data on a scale which is large enough to show convincing results. This difficulty is highlighted by (Coughlin, 2003),

“...human evaluation has serious drawbacks: in addition to relying on subjective judgments, it is both time-consuming and costly. This in turn means that the scale of these evaluations tends to be so small – usually no more than a few hundred sentences examined by a small number of raters – that it can be difficult to draw firm conclusions about system quality.”

This problem may have become exacerbated by the fact that (SMT) Statistical Machine Translation has become a more dominant approach to MT than rule-based MT (RBMT) over the past few years[17]. This dominance is evidenced, for example, by RBMT system participation rates in MT evaluation competitions. In SMT it is more difficult to focus on a particular linguistic phenomenon. A larger volume of data may be required to evaluate convincingly a broad-coverage SMT system than would be necessary if, for example, an RBMT system which focuses on adjective endings in German were being evaluated by volunteers using a “glass box” (Olive, Christianson, & McCary, 2011 p. 745) testing approach.

Paying experts or non-experts

Non-experts

A common alternative to volunteers is to pay local non-experts (usually students) or more expensive experts (usually professional translators) as a means of gathering manual evaluation data, for example, (Coughlin, 2003). The availability of bi- or multilingual students in learning institutions in which research into MT is being carried out makes the logistics of data elicitation easier to manage and certain experimental variables easier to control.

Crowdsourcing

A recent innovation in the field of MT evaluation is the use of “crowdsourcing[18]” to reduce the cost of manual data elicitation from non-experts.

Statistics for data collected on MTurk[19] for the ranking task in total, 55,082 rank labels were collected across the eight language pairs (145% of expert data). Each language pair had 600 sets, and we requested each set completed by 5 different workers. ... The total cost of this data collection effort was roughly $200. (Callison-Burch et al., 2009 p. 19)

This non-expert approach has been shown to correlate well with scores elicited from experts during a WMT shared task MT evaluation when ranking MT systems[20] but it suffers from two disadvantages:

Intuitively, different language pairs may have different associated costs as the cost-of-living in the countries where the target language is natively spoken varies. Thus, for example, we would expect English to Swedish to cost more than English to Urdu.

The cost of data acquisition is still linear to the quantity of evaluation data being gathered so the cost saving, though considerable, is only one of degree.

Despite its disadvantages we believe crowdsourcing is a promising lower-cost approach to MT evaluation, particularly when combined with unobtrusive methods.

Experts

Due to budgeting constraints the use of experts to elicit MT evaluation data is not as common a practice as non-expert elicitation. Also, the term *expert* can be ambiguous. When we refer to experts in the context of PEMT in an unqualified way, we mean trained professional translators whose primary income is derived from the act of translation.[21]

Unobtrusive methods for MT evaluation

In general, MT evaluation tasks can be split into two categories: labelling (sometimes called "annotating") and post-editing. As labelling is not a routine task for a translator, it is considered by us to be *obtrusive* to the task of translation. It would cost money to have a professional translator carry it out as part of a commercial translation project. By contrast, post-editing is an *unobtrusive* method or task; i.e., one which, ipso facto, must be carried out in the context of PEMT.

While post-editing string data are informative with regard to MT evaluation, other data which can be gathered by means of instrumentation may also be of value. For example,

Temporal data

For example, how much time does the translator spend post-editing each segment?

Behaviour patterns

For example, was the MT proposal discarded by the post-editor and, if so, after how long? How does the translator interact with the CAT tool environment when carrying out a PEMT task?

Mouse and keyboard interaction

For example, what are the pause and typing durations?

Interactions with CAT tool functions
For example, concordancer use, tag reordering, use of a termbase etc.

Similar PEMT studies

In general, wherever measurement of post-editing is used to evaluate the quality of raw MT we consider the method to be unobtrusive. In addition, we are concerned with post-editing combined with various software instrumentation methods (for example, to measure translator speed).

A non-commercial instrumented PEMT study

Although our approach is generally designed to be applied to commercial PEMT projects where translators are being paid, a similar approach to ours is described in a non-commercial study of post-editing speed by (Tatsumi, 2009).

The study involved an analysis of 1413 sentences to measure the correlation between automatic evaluation metric scores[22] and post-editing speed using an instrumented CAT tool for three professional translators.

We believe this research could be extended to measure the same correlation for different languages and also to test other string comparison methods. This is an interesting problem, because a one-size-fits-all approach that uses, for example, the BLEU metric may not be appropriate for certain language pairs. This point is illustrated by (Ananthakrishnan et al., 2007), who present a study that shows that simplistic n-gram matching cannot differentiate between certain features of Hindi. This has further implications regarding the use of specific automated metrics where temporal data is not available (for example, evaluation of an MT engine for a specific language using a corpus of raw and post-edited sentences). In general, an automated MT evaluation metric which correlates well with manual post-editing effort (measured by time) for a particular language or text type is better than one that does not when we are concerned with PEMT.

Two commercial PEMT productivity studies

Groves and Schmidtke (2009) use an un-instrumented approach in Microsoft to compare raw MT output with its human post-edited counterpart to identify post-editing patterns which could be used to improve post-editor productivity. They also present data which illustrates increased productivity (varying from 6.1% for English to Czech to 28.6% for English to Danish) for PEMT versus a HT process. As this study uses commercial post-editing data, its sample sentence count is much larger that Tatsumi, for example, just under 10,000 sentences for German to English alone. Regarding the need for project specific MT suitability studies, Groves & Schmidtke have the following to say:

“Considerable variation between similar languages such as Swedish and Danish for this specific project is more likely caused by factors at the translation stage rather than significant core MT system quality differences between the languages. This illustrates

why productivity measurement is necessary to assess suitability of MT for a specific project and why standard MT evaluation metrics are not always reliable."
(ibid. p. 2)

A second commercial study of PEMT productivity was carried out in AutoDesk Inc.[23] (Plitt and Masselot, 2010). They compared PEMT and HT productivity over a two day translation task. We believe their instrumented approach most closely mirrors our proposed approach. They used an open-source toolkit called Moses[24] (Koehn et al., 2007) which was used to create an SMT system. It was trained on company-internal translation memories. They measured the PEMT speed of 12 individual translators spread evenly across four languages. In total 144,648 source words were analysed.

Plitt & Masselot's instrumented approach is similar to that of Tatsumi's in that they measure individual translator speed on a segment-by-segment basis. Their results show significant PEMT speed variance across individual translators. This ranged from +20% to +131%, where 100% represented a doubling of translation productivity. We regard the identification of well-motivated[25], fast and good MT post-editors as being of significant commercial relevance for PEMT.

Regarding normalisation of Plitt and Masselot's results for quality, they report the following:

"Our expectation was that the quality of the post-edited translation would be equivalent to traditional translation, quality being defined here according to the standard criteria applied at Autodesk. To verify that this expectation was met, we provided the Autodesk translation QA team with samples of translated and post-edited text, again randomly selected, and of reasonable size. The QA team was aware of the overall context of the productivity"
(ibid. p. 10)

Their results showed that PEMT text samples flagged significantly fewer translation errors than manually translated text. This was remarked on by the authors as a "surprising" result. An interesting extension of this study to confirm this result would be to see if QA reviewers could guess which text samples were PEMT and which were HT. Based on their description of the QA process it is theoretically possible that reviewers were able to tell which sections were PEMT versus HT and left the PEMT sections unchanged but imperfect because they did not know where to start carrying out corrections. A guessing game test would rule this out.

This study is particularly interesting because it shows how an open-source SMT toolkit which was made available through a publicly-funded research effort could be used to produce a measureable increase in translator productivity in a commercial context. This is relevant to our vision that, under the right circumstances, commercial post-editing data can be of some use in comparative MT research.

For example, Plitt and Masselot's study could be repeated to test an improvement in the Moses decoder component, using the data from the first study as a baseline. If this repeated study found an improvement the manually evaluated results could be published and this result could itself be used as the next baseline, and so on. Our "operationalised" approach is intended as a means of reducing the effort involved in repeating this experiment in a translation project. A closed-source or proprietary system could also supply the MT output but the impetus to publish results of this experiment would be reduced as owners of commercial systems do not normally wish to publish detailed studies on how they achieve MT better results, lest competitors seek to duplicate their success. Also, our system may make it easier to judge the return-in-investment for proprietary systems versus freely available open-source ones[26].

Of less interest to the research community but of some practical concern for those involved in the training of SMT systems for commercial projects is the choice of

Our approach

The system we present below overlaps with Plitt and Masselot's study in the following ways:

We wish to gather MT post-editing data in an unobtrusive manner in a commercial translation context to answer the question of whether PEMT is desirable or not relative to HT within a specific translation project.

It uses an instrumented approach to gather data beyond post-editing string differences (for example, time data).

However, our approach differs in a number of respects:

In addition to comparing HT to PEMT we seek to compare MT with MT. Our approach is designed to evaluate MT suitability in general for a large translation project and then to compare MT systems with each other, once a suitable baseline has been established by means of access to an MT brokering system.

We wish to be able to apply our approach in a variety of translation project settings. To this end, we seek to operationalise our approach by connecting our system to a TMS which can keep track of operational variables like translators, TM's, word rates, and translation jobs. We hope that by maintaining a loose coupling via web services to a TMS system, we can more easily apply our approach to new large translation projects that keep operational data stored in a TMS. The aim is to make it possible to carry out similar studies to Plitt and Masselot's with minimal cost overhead.

We believe that results should be gathered in an instrumented CAT tool. We believe that results gathered using a CAT tool in a manner similar to Tatsumi's approach are more authentic, as most translators make heavy use of CAT tools at their desk. This observation has been confirmed in a study carried out by (Lagoudaki, 2006) who found that 82.5% of translators who responded to her survey used a CAT tool.[27] Our initial approach has been to adapt an existing open-source CAT tool called OmegaT but we aim to keep our methods as generic as possible. To that end, we will also attempt to add some instrumentation functionality to Trados, the best-known CAT tool in the technical translation industry, according to (Lagoudaki, 2006 p. 16). We also aim to gather data on MT suggestion discard rates, where translators feel that it is quicker to delete an MT suggestion than post-edit it.

Motivation for our choices of existing software components

In this section we will elucidate the reasons for our approach and list some of the reasons we have chosen to use certain software systems in our research.

Motivation for comparing MT with MT as well as PEMT with HT

We believe that once PEMT has been established as appropriate for a large or ongoing translation project[28] the natural follow-on question is whether any other MT system or training corpus can be used to improve on this baseline. We also believe that, despite inherent intellectual property restrictions, manual evaluation results garnered for experimental MT systems in commercial translation projects are valid and publishable in a research context. For this reason it should be made as easy as possible to repeat studies similar to that of (Plitt and Masselot, 2010) to test MT systems (engines and training corpora) against each other.

Motivation for connecting our system to a TMS

The following paragraph describes some of the variables which affect PEMT productivity in Microsoft:

"There can be substantial variations in post-editing productivity, for the same language, between different projects and products, different handoffs during the same project, and different translators. This indicates that MT language quality itself is only one of several important factors influencing productivity. How closely the text to be MT'd correlates with training data is an obvious quality factor. Certain types of text work better for MT than others, and a formal writing style helps, although this may be more a matter of what text types are most common in training data than of inherent issues with MT itself." (Groves & Schmidtke, 2009 p. 2)

This long list of factors that influence PEMT productivity in Microsoft, an international software vendor, can also apply to language service providers and indeed individual translators who may deal with different text types every day for different languages and different clients. In addition, client and end-user quality expectations may vary, so that, on the low end of the quality expectation spectrum, fidelity (informational accuracy) may be judged to be sufficient. On the high end, a client may insist that the final translation product should be stylistically indistinguishable from texts manually translated and reviewed by experienced professional translators. Finally, once MT has been judged appropriate by all stakeholders for a particular project, the ongoing question remains as to whether the MT system used in production can be improved upon. In the case of statistical MT engines, a particular training corpus may indeed yield better results. Equally an algorithmic improvement may also lead to better performance.

These factors combine to make ongoing tracking of MT post-editing in a commercial context a complicated and time consuming task. Thus, any system whose aim is tracking post-editing effort in a multi-client, multi-project commercial context in a way that does not create an explosion in organisational effort at the project management level should be able to access an existing translation management system which already contains information pertaining to entities like clients, files, language pairs, word rates, translators and reviewers.

Motivation for choosing GlobalSight as our TMS

A number of factors combine to make GlobalSight our TMS of choice.

- It is free and open-source.
- It is an active project which is undergoing continual improvements.
- It has a sophisticated web services application programming interface (API).
- It is functionally similar to existing proprietary solutions like SDL's WorldServer[29].
- It is owned by WeLocalize[30], who are a CNGL[31] industrial partner.

Motivation for connecting our system to a CAT tool

The incidence of CAT tool usage amongst translators is outlined in a 2006 survey (Lagoudaki, 2006), that found that 82.5% of respondents (who were mainly technical translators) used a CAT tool. The importance of CAT tools for translators is highlighted by Somers (Somers, 2003 p. 31), who states:

"one of the most significant computer-based aids for translation is the now widely used translation memory (TM)."

The advantages of CAT tools for individual translators are complemented by their usefulness in large multi-translator translation projects, in particular with regard to terminological consistency, which is one of the more challenging aspects of quality control for large translation projects.

Even where translators are located physically in the same room, maintaining terminological consistency is difficult. When translators who have never met in person work in parallel on a project, this challenge is magnified. CAT tools have important functions to aid in maintaining terminological consistency, which would make their omission in a contemporary multiple-translator translation project almost unthinkable. These functions are:

Terminology management

Because of the cost overhead involved in its compilation, an extensive terminology database may not always be deemed financially feasible for an individual translation project, but when it is available, a terminology database along with term matching in a CAT tool can play an important role in maintaining terminological consistency. Some tools also provide a mechanism for running a semi-automatic quality analysis using the terminology database (for example, the “QA Tool” function in Trados). This works in batch mode and checks that if a source segment contains a term found in the terminology database, the translation of that term appears in the target segment, though its utility can often be limited by a high incidence of false positives, indicating terminological inconsistencies which are desirable.

Concordancing

Even if working without using an online translation memory, a regular merge of translation memories from translators working in parallel can make it possible for one translator to see how another project colleague translated a term in the past by searching the project translation memory for a word or phrase using the CAT tool concordance function. As it may be a frequently accessed function, it should be as easy to access as possible while translating.

Full and partial segment matching

This can be measured at the outset of a project. It plays an important role in the economics of any translation project where text is repeated. This is particularly true for projects which are an update of previously completed projects.

The PEMT quality spectrum

At the positive extreme, machine translated text which is reviewed but left unchanged by a post-editor is implicitly described by the post-editor as fit for purpose, and the argument for using MT in the translation production cycle at hand is clearly bolstered. Conversely, if the same target text were only comprised of sentences which must be completely rewritten, the argument for using MT in this translation process is negated, as the time required to read, mentally parse, reject an MT suggestion outright and then translate is generally greater than the time required simply to translate the sentence without the aid of MT. Thus, we can say that in theory, irrespective of quality concerns, MT may reduce the number of words a translator can translate per unit of time, which is most commonly the primary productivity measure for translators. In practice, we expect most translations using MT to tread a middle ground between these two extremes. While we accept that methods[32] exist to measure MT post-editing using commonly available CAT tools like Trados or Omega-T, a number of weaknesses exist with regard to this approach. In general these weaknesses arise from using common commercial and open-source CAT tools not designed to measure post-editing behaviour. In particular, a number of questions related to post-editing cannot easily be answered using current CAT tools.

How much of the segment was changed?
Below a 50% threshold, using the Levenshtein (Levenshtein, 1966) string distance algorithm, we cannot tell if any of the raw MT was kept by a post-editor.

How long did each segment take to post-edit or translate?
Although we can estimate times based on daily productivity for a single translator, this does not tell us anything about individual sentences. For example, it may be quicker to translate two sentences than one complicated long one. For a description of how controlling the source language can help speed up PEMT see (Aikawa et al., 2007).

How long did it take the translator to decide the MT suggestion should be discarded completely?

A post-editor may discard an MT proposal immediately after entering each segment or repeatedly so we could infer that she has lost faith in the MT system's ability to produce salvageable proposals for this file.

CAT Instrumentation

Instrumentation of software is described by (Kim et al., 2008) as the automatic recording of user behaviour within a system. In general, it is technically difficult or impossible to add arbitrary instrumentation to commercial software which is disseminated in binary form. Binary distributions of the application are hard or impossible to adapt, since their source code is not available. Where an application programming interface or set of web services to an application does exist, it may not expose the internal functions needed to carry out the instrumentation and it cannot be used to make arbitrary changes to the product's graphical interface. Fortunately, in recent years free open-source CAT tools like Anaphraseus[33], OpenTM2[34] and OmegaT have become available. The open-source CAT tools chosen as the clients for the system described is OmegaT.

Advantages of OmegaT relative to other open-source CAT tools

It has an active developer and user community. As long as it remains so, support and gradual improvements can be expected.

It is performant. Preliminary tests on a large translation memory containing 326,712 segments showed 100%, fuzzy matching and concordancing speeds subjectively similar to those of Trados, the system used as the performance benchmark.

It can process a number of common file formats. The application has a large set of robust file filters which can separate translatable text from formatting code in different file types like HTML and Microsoft Office™ documents.

It is functional. Compared to Trados, which was also used as a functional benchmark, it contains most of the functionality that translators commonly use in a CAT tool, for example, fast concordancing, terminology viewers with terminology matching, interfaces to online MT engines and fuzzy matching.

It is easy to disseminate.

The application is quick and uncomplicated to install via a Java web start (JNLP) link, which should aid dissemination.

Disadvantages of OmegaT relative to Trados

Lacking in some important features
OmegaT lacks some features found in the functional benchmark, Trados; for example, tag locking (which makes formatting tags read-only) and variable tag visibility are not present.

Lacking in some secondary features
For example, no batch mode terminology QA.

More limited file format support
OmegaT generates a large number of placeholder tags for Microsoft Office file formats (for example, docx), which combined with a lack of variable tag visibility function can make it hard for translators to distinguish translatable from non-translatable text.

A more limited user base
Although an open-source CAT client has obvious benefits for comparative research as well as functional and graphical user interface adaptability, OmegaT is only used by a small minority of translators and an even smaller minority of translation agencies. Based on Lagoudaki's CAT tool usage statistics (Lagoudaki, 2006) and experience in the past with large scale commercial translation projects, we are aware that most commercial translation workflows assume Trados as the CAT client. For this reason we will attempt a parallel prototype development strategy using the Trados application programming interface.

System Design

The working title we have chosen for the whole system is "try-and-see-mt", to underline its agnostic position with regard to PEMT versus HT. We will use this name when we refer to the system as a whole. The TMS and CAT components will be adapted to work with the try-and-see-mt system, but they can be considered existing components. All inter-component communication will use web services.

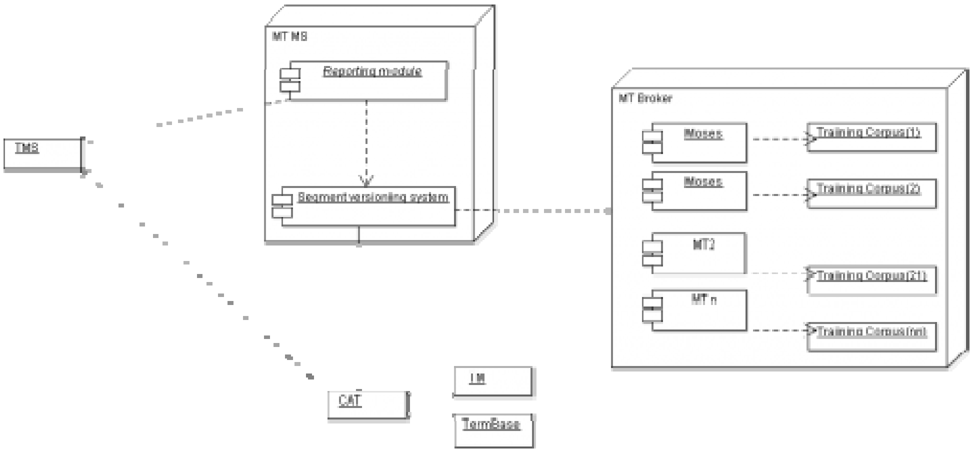

Fig. 2: UML deployment diagram to describe the try-and-see-mt component architecture

In general the system has three modes of operation, depending on the evaluation being carried out. Each mode is just a generalisation of the mode preceding it, so that at a technical level as little distinction as possible will be made between them.

- HT versus PEMT
- MT versus MT
- Multiple competing MT systems

System components

The instrumented CAT client (CAT)

As our system is designed to compare multiple competing MT systems, we will use a look-ahead and caching mechanism in the CAT client to fetch translations from the MTMS and save them in a temporary cache on the translator's hard drive. Each fetch operation will be carried out by a low priority Java thread. The decision as to how many threads will be allowed to access the MTMS will be made during development.

This fetch mechanism can be thought of as a cursor which moves from the first to the last segment in a file initiating a number of web service requests to the MTMS and caching the responses. Particularly, in multiple competing MT systems mode, the CAT tool may not cache too many responses as this would degrade the MTMS's ability to adapt by omitting poorly performing MT systems from the pool of MT candidates.

When the translator is finished translating the file, she can confirm this by clicking on a “Finish File" button in the CAT tool. This will send a message to the MTMS that the file is finished, which will make the segments associated with that file read-only unless the translator decides to unlock them manually. The translator can confirm that the project is finished by clicking a “Finish Project” button in the CAT tool. At this point segments will be read-only unless they are unlocked on the MTMS server (for example, by a project manager or system administrator). This is equivalent to the translator's delivering the final files to the client by email. All translated text will be stored locally and on the MTMS component, which will fulfil a secondary function as a redundant backup to protect against data loss. Note, this description applies to OmegaT only.

The Machine Translation Management System (MTMS)

Initially this component is a container for the SVS and Reporting Module.

The Segment Versioning System (SVS)

The Segment Versioning System (SVS) keeps track of all changes to segments along with who made them and when and the providence of MT proposals which are accessed via the MT broker. All changes to the segment will be logged when a segment is closed (i.e., when a translator moves from one segment to the next) but for reporting purposes our main concern will be the string value for the segment when the "Finish project" flag is set to true. Segments will be uploaded using a low priority thread which checks that a local file is fully synchronised with the MTMS when the "Finish File" or "Finish Project" button is clicked in the CAT client. The SVS will also store data from the CAT tool instrumentation module such as time data and whether MT proposals were discarded by the translator carrying out the post-editing (using a keyboard shortcut or button in the CAT client).

The reporting module (RM)

The reporting module will collate all the post-editing, time and MT suggestion rejection data along with the information stored in the TMS such as project identfiers and translator identifiers. This information will be provided to the project manager in the form of a report.

Some system design constraints

Translators can continue to translate despite an intermittent Internet connection This constraint is motivated by the fact that translators, in particular freelance translators, are most commonly paid on a per word basis. Any delay caused by an intermittent network connection or slow MT provisioning service results in reduced earnings. The system should not make the translator wait for an MT service which may not be of use. If an MT provisioning server (for example, an MT broker server similar to that described in (Federmann and Eisele, 2010) is not available, it should not prevent a translator from working with the CAT tool, even if this means that no MT is provided for a sentence.

Translators can still deliver their translations if the system is not available
This constraint is motivated by the fact that translation jobs are handed off and handed back on tight schedules. A system which fulfils a non-mission-critical role like the try-and-see-mt system should not increase the risk of a late translation job because of a technical failure.

The system can adapt to omit poorly performing MT systems from the list of candidate MT systems
If one of the MT systems connected to the MT broker is performing demonstrably worse than its fellow candidate systems as measured by post-editing string distances and sentence discard rates, we may wish to stop using this MT system early in the translation process. This means that we cannot batch machine translate large volumes of text using the TMS system, since we do not know which MT systems will be omitted as candidates during the translation process. In general, we wish the MTMS to process as few sentences as possible in its attempt to find the best performing MT system. Candidates that perform similarly will require a larger sample set of sentences to determine statistically significant performance differences so poor candidates should be removed early on.

The system can be disconnected from the translation workflow with ease
In general, once a single MT source has been identified as good, it should be possible to decouple the try-and-see-mt system from the translation workflow with as little effort and interruption for the project participants as possible.

Next steps

We plan to carry out a number of experiments to test our approach. In particular,

We wish to measure the correlation between post-editing distances when sentences are post-edited by a single translator on a translation team and when sentences are post-edited by many or all translators.

We would like to test our unobtrusive approach using obtrusive labeling techniques via Amazon's Mechanical Turk crowdsourcing platform.

We would like to compare various QA techniques (e.g. error categorisation) in comparrison with measuring post-editing carried out by a second translator (acting as a reviewer).

Summary

As commercial use of PEMT use increases so too does the need to compare MT system output. We hope that by providing an instrumented platform to measure post-editing in a project setting our work may be of use to other researchers working in the area of post-editing, MT development and MT evaluation.
The system outlined in this paper is currently under development in the Centre for Next Generation Localization; its development can be tracked at the website www.try-and-see-mt.org.

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin. The authors would like to thank Dr. Nancy Schmeing for helpful feedback on an early draft of this paper.

Bibliography

Aikawa, T., L. Schwartz, R. King, M. Corston-Oliver, and C. Lozano (2007), Impact of controlled language on translation quality and post-editing in a statistical machine translation environment, in Proceedings of MT Summit XI, pp. 1-7, Copenhagen, Denmark.

Allen, J. (2004), Case Study: Implementing MT for the Translation of Pre-sales Marketing and Post-sales Software Deployment Documentation at Mycom International, in Machine Translation: From Real Users to Research, vol. 3265, edited by R. Frederking and K. Taylor, pp. 1-6, Springer Berlin / Heidelberg.

Ananthakrishnan, R., B. Pushpak, and M. Sasikumar (2007), Some issues in automatic evaluation of english-hindi mt: more blues for bleu, in In proceeding of 5th International Conference on Natural Language Processing, vol. In proceed, pp. 135-139, Hyderabad, India.

Banerjee, S., and A. Lavie (2005), METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pp. 65-72, ACL, Michigan, USA.

Callison-Burch, C., P. Koehn, and C. Monz (2009), Findings of the 2009 workshop on statistical machine translation, Machine Translation, (March), 1-28.

Callison-Burch, C., P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. F. Zaidan (2010), Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation, in Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, p. 17–53, ACL.

Coughlin, D. (2003), Correlating Automated and Human Assessments of Machine Translation Quality, in Proceedings of MT Summit IX, pp. 23-27, New Orleans, USA.

Doherty, S., and S. O'Brien (2009), Can MT Output Be Evaluated Through Eye Tracking?, in Proceedings of MT Summit XII, pp. 214-221, Ontario, Canada.

Federmann, C., and A. Eisele (2010), MT Server Land: An Open-Source MT Architecture, The Prague Bulletin of Mathematical Linguistics, 94, 57–66.

Groves, D., and D. Schmidtke (2009), Identification and Analysis of Post-Editing Patterns for MT, in Proceedings of MT Summit XII, pp. 429-436, Ontario, Canada.

Guerberof, A. (2009), Productivity and Quality in MT Post-editing, in Proceedings of MT Summit XII, Ontario, Canada.

Kim, J. H., D. V. Gunn, E. Schuh, B. Phillips, R. J. Pagulayan, and D. Wixon (2008), Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems, in Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, p. 443–452, ACM.

Koehn, P., H. Hoang, and A. Birch (2007), Moses: Open source toolkit for statistical machine translation, in Proceeding ACL '07 Proceedings of the 45th Annual Meeting of the ACL, pp. 177-180, ACL.

Krings, H. (2001), Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes, The Kent State University Press.

Lagoudaki, E. (2006), Translation memories survey 2006: User's perceptions around tm use, in Proceedings of the ASLIB International Conference, vol. 2006, pp. 1-29, London, UK.

Levenshtein, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady, 10(8), 707-710.

Li, Z., C. Callison-burch, C. Dyer, J. Ganitkevitch, A. Irvine, L. Schwartz, W. N. G. Thornton, Z. Wang, J. Weese, and O. F. Zaidan (2010), Joshua 2 . 0: A Toolkit for Parsing-Based Machine Translation with Syntax, Semirings, Discriminative Training and Other Goodies, Machine Translation, 2-6.

Olive, J., C. Christianson, and J. McCary (Eds.) (2011), Handbook of Natural Language Processing and Machine Translation, Springer, New York, USA.

O'Brien, S. (2005), Methodologies for measuring the correlations between post-editing effort and machine translatability, Machine Translation, 19(1), 37–58.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002), BLEU: a Method for Automatic Evaluation of Machine Translation, Computational Linguistics, (July), 311-318.

Plitt, M., and F. Masselot (2010), A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context, The Prague Bulletin of Mathematical Linguistics, (93), 7-16.

Somers, H. L. (2003), Computers and translation: a translator's guide.

Tatsumi, M. (2009), Correlation between automatic evaluation metric scores, post-editing speed, and some other factors, Proceedings of MT Summit XII, (2001), 332-339.

Notes

[1] http://www.sdl.com (a large translation software and services company)

[2] http://www.sdl.com/en/language-technology/landing-pages/machine-translation-survey

[3] http://www.translationautomation.com

[4] We acknowledge that large companies that use PEMT may have efficient methods for measuring PEMT productivity but, if they exist, they are not made available outside of the company in question.

[5] This volume versus granularity dichotomy is common in all forms of research

[6] http:///www.omegat.org

[7] http://www.trados.com

[8] see for a concise summary of a number of automated MT metrics

[9] Other variants to this approach exist, for example, using many manually or automatically generated reference sentence versions, but studies of this nature are less common so we do not discuss them here.

[10] Papineni et al stress this fact further by stating in their first footnote that the letter 'U' in the acronym "BLEU" stands for "understudy", which means a subordinate replacement actor who can stand in for a colleague in theatre.

[11] "Reinhard Schäler interviews Andy Way on MT", transcribed from http://www.youtube.com/watch?v=pgg7_A0Rla8, uploaded Sept. 2009. Permission to cite was sought and granted by Andy Way.

[12] A term coined by Jeff Howe, http://www.wired.com/wired/archive/14.06/crowds.html

[13] http://www.statmt.org

[14] http://iwslt2010.fbk.eu

[15] The most common commercial text type for post-edited machine translation (PEMT) is technical translation (see http://www.sdl.com/en/language-technology/landing-pages/machine-translation-survey), but news is an obvious candidate genre for gisting, and, in general, it does not carry the same intellectual property restrictions as technical translation, which may contain sensitive product information.

[16] In practice, it takes considerable time to produce improvements in MT systems, so their frequency is not as big an obstacle to progress as might be supposed.

[17] In the findings of the WMT 2010 (ibid.) it was considered unfortunate that fewer RBMT systems took part. However, RBMT systems have been shown to compete well

with SMT systems for closely related languages, so perhaps a language pair from this category might entice competitors from this camp to take part.

[18] http://en.wikipedia.org/wiki/Crowdsourcing

[19] https://www.mturk.com/mturk/welcome

[20] However, intra-annotator agreement is reported as low between groups of (MT system development) experts and crowd-sourced annotators.

[21] In the context of the WMT competitions, experts are researchers in the field of MT taking part in the shared task.

[22] The results showed that most of the metrics were quite close in terms of correlation with time data. An interesting extension would be to see if the Levenstein distance also correlates well, as this is the string comparison method used in most CAT tools.

[23] http://www.autodesk.com

[24] http://www.statmt.org/moses

[25] We suspect that translators who only carry out PEMT may become bored over time, but, based on anecdotal evidence, translators who only carry it out for a day or two per week may welcome the change of modality on the basis that "a change is as good as a rest".

[26] We acknowledge that propriety MT systems often need considerable investment to optimise their output but we believe that this will become less over time.

[27] Most of the respondents who responded to her survey described the translation work they most frequently carried out to be of a "technical" nature. This coincides with the results of a commercial survey that found that PEMT was most often applied to text of a technical nature.

[28] A project is considered by us to be comprised of a translation need, client or end-user quality expectations and a pool of language professionals involved in the production of the final translation product.

[29] http://www.globalsight.com/wiki/index.php/Comparing_GlobalSight_with_WorldServer

[30] http://www.welocalize.com

[31] http://www.cngl.ie (Centre for Next Generation Localization)

32 See http://www.try-and-see-mt.org/html/existing-methods-to-measure-mt-post-editing.html for a description of a method to use a standard CAT tool (in this case, Trados or OmegaT) to measure post-editing across a batch of files.

33 http://anaphraseus.sourceforge.net/

34 http://www.opentm2.org (formerly IBM Translation Manager)