Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost

Lingxiao Wang **Christian Boitet**
UJF, LIG-GETALP, BP 53
41 rue des Mathématiques, Domaine Universitaire
38041 Grenoble Cedex 9
Lingxiao.Wang@imag.fr, Christian.Boitet@imag.fr

Abstract

An interactive Multilingual Access Gateway (iMAG) dedicated to a website S (iMAG-S) is a good tool to make S accessible in many languages immediately and without editorial responsibility. Visitors of S as well as paid or unpaid post-editors and moderators contribute to the continuous and incremental improvement of the most important textual segments, and eventually of all. In this approach, pre-translations are produced by one or more free machine translation (MT) systems. Continuous use since 2008 on many websites and for several access languages shows that a quality comparable to that of a first draft by junior professional translators is obtained in about 40% of the (human) time, sometimes less. There are two interesting side effects obtainable without any added cost: iMAGs can be used to produce high-quality parallel corpora, and to set up a permanent task-based evaluation of one or more MT systems.

1 Introduction

An iMAG is an interactive Multilingual Access Gateway very much like Google Translate at first sight: one gives it a URL and an access language and then navigates in that access language. When the cursor hovers over a segment, a palette shows the source segment and proposes to contribute by correcting the target segment, in effect post-editing a MT result or improving on a previous post-edition. With Google Translate, the page does not change after contribution, and if another page contains the same segment, its translation is still the rough MT result, not the polished post-edited version. The more recent Google Translation Toolkit enables one to MT-translate and then post-edit online full web pages from sites such as Wikipedia, but again the corrected segments don't appear when one later browses the same page in the access language.

By contrast, an iMAG-S is *dedicated* to an *elected website* S, or rather to the *sublanguage* empirically defined by the textual content of one or more URLs constituting S. The iMAG-S contains a translation memory (TM) and if possible a specific, pre-terminological dictionary (pTD) (Daoud et al., 2009), both dedicated to the elected sublanguage. Segments are pre-translated not by a unique MT system, but by a (selectable) set of MT systems. Systran, Reverso and Google Translate have been mainly used as well as Neon for Chinese-English, but specialized systems developed from the post-edited part of the TM, and based on Moses (Koehn et al., 2007), are also used in our gateway.

The online contributive platforms SECTra_w (Huynh et al., 2008) and PIVAX (Nguyen et al., 2007) are used to support the TMs and pTDs. Translated pages are built with the best segment translations available so far. While reading a translated page, it is possible not only to directly post-edit the segment under the cursor, but also to seamlessly switch to SECTra_w online post-editing (PE) environment, equipped with filtering and search-and-replace functions, and then to go back to the reading context. To illustrate our points, we will use an iMAG created for the website of our lab (400 researchers, 25 teams).

Since 2008, we have regularly added iMAGs to our platform, and found that two interesting side effects are obtainable without any added cost: iMAGs can be used to produce high-quality parallel corpora, and to set up a permanent task-based evaluation of one or more MT systems.

Sharon O'Brien, Michel Simard and Lucia Specia (eds.)
Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, September 2, 2013, p. 103–110.

Figure 1: Access in Chinese of to the LIG lab website

2 Typical scenario of use

2.1 Multilingual access to a website

Figure 1 shows the iMAG access interface to the LIG lab website. We choose Chinese as the access language from the pull-down menu. One or more free MT servers, in this case Google Translate and Systran, produce initial translations.

2.2 Post-editing and scoring on the page

As shown in Figure 2, when the mouse pointer hovers on a segment (title, sentence, menu item), an interactive palette pops up. It's dialogue box displays the source language content (in blue), and users can post-edit and evaluate the text in the access language.

Figure 2. Direct PE of translation results

All visitors of the page can contribute by post-editing. However, only registered users can use the "Advanced mode". If the TM contains several post-editions for one segment, the system selects which to use in the translation page, based on highest score and then on most recent time.

2.3 Visualization of translation reliability

In the translated page, users can see the estimated *reliability* of each segment by checking the "Reliability" checkbox. As shown in Figure 3, colored brackets {_…_} enclose each translated segment. If a user post-edits this page, colors[1] change based on the user's profile[2]. If one clicks the "Original" button (in the upper right corner of figure 3), the left side of the browser window displays the page in the access language, and the right side the original page.

2.4 Post-editing TMs in "Advanced Mode"

The "Advanced Mode" offers a translation editor interface similar to those of translation aids and commercial MT systems, that makes post-editing much faster than in the presentation context. Not yet post-edited segments can be selected, and global search-and-replace is available. Figure 4 shows a screenshot of SECTra_w PE interface.

When an iMAG-S is created, we select several MT systems for proposing *pretranslations,* and set the *preferred* one. That can be changed later. From the post-editing interface, it is possible to perform various operations on the TM:

- MT results: discard a MT result, call again one of the selectable MT systems, and use an MT result as current post-edition[3].
- Post-editions: discard a post-edition, use one as preferred in the current context.

[1] Green: privileged users; Orange: anonymous users; Red: MT output (the translation results have never been edited).
[2] For that, the page must be refreshed.
[3] That result is then moved to the PE cell.

Figure 3. iMAG page display in “Reliability” + “Original” mode

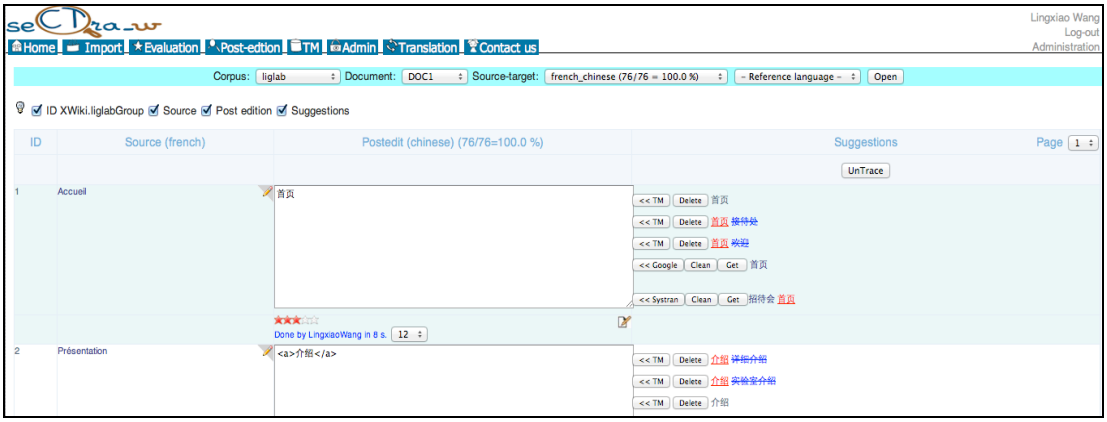

Figure 4. Advanced mode (SECTra_w screenshot)

As shown in Figure 4, users can visualize and compare the edit distances between the chosen current translation and the MT and PE results for that segment, contained in the translation memory, using the "Trace/Untrace"[4] button.

2.5 User profiles, moderation, scoring

The admin assigns a profile to each registered user: *reliability* (bilingual, professional translator, translator certified for site S) and *quality score* for each language pair (from 0 to 20). That is based on language tests (A1-…C2, or ILTS, TOEFL, TOEIC, etc.), or on professional levels. *Moderators* are contributors competent enough in the domain of S and ‘blessed’ by S.
The *quality score* of a post-edited segment is, by default, that of its post-editor. It may be changed by the post-editor herself (self-evaluation), and also later by an admin or by a moderator.[5]

[4] A "mixed character/word edit distance" is used.
[5] There is also a subjective evaluation environment, where several judges can participate, assigning classical scores.

3 Conclusions after 4 years of use

After the first four years of use, there are over 80 demo iMAGs in operation, sharing 3 TMs, and 6 dedicated iMAGs, with their own TMs. There are 8 source languages, and websites can be accessed in more than 10 languages with post-editing support. There are more than 820,000 segments, and about 45% (370,000 +) segments have been post-edited by contributors. Most parallel segments are English-French, English-Chinese, and Chinese-French. We give some statistics in Table 1.

3.1 Reliability and quality indicators

It is quite difficult in practice to maintain a website in more than one language. Take for example the website of Figure 1, which an admin tries to maintain also in English: hitting the “English” button directs to web pages that are still 85% in French, while the 15% portion in English is far from perfect. By contrast, using the iMAG button and choosing English shows pages 100% in Eng-

lish, and the reliability can be shown for each segment. The reliability and score of each segment are also visible in advanced mode. More important, the overall quality of the post-edited segments (green or orange special brackets if shown), estimated by teachers of English and bilinguals, is at least as good as that of translations (when available) found in the static "English version".

Language pair (L1→L2)	Bi-segments	Source Words L1 (Standard p.)	Target Words L2 (Standard p.)	Size L1	Size L2
English → French	121 074	2 542 731 (10 170 p.)	2 613 351 (10 453 p.)	10,1MB	10,4MB
English → Chinese	208 106	4 370 530 (17 482 p.)	6 063 942 (151 159 p.)	19,1MB	17,6MB
French → English	29 079	627 661 (2510 p.)	610 098 (2 440 p.)	4MB	3,9MB
French → Chinese	10 890	228 703 (914 p.)	317 322 (793 p.)	1,5MB	1,25MB
Chinese → English	2 013	58 656 (146 p.)	42 275 (169 p.)	240KB	263KB
Chinese → French	10 062	291 192 (727 p.)	211 185 (844 p.)	874KB	1MB

Table 1. Parallel segments obtained (we count the number of Chinese characters for Chinese) (250 words/page in English/French, or 400 Chinese characters/page.)

The "trick" behind this is simple: no target pages are kept anywhere in our system. Only individual segments are kept (in the dedicated TM). Pages in target languages are dynamically built using the best translation (MT output or PE) available for each segment. We decide which string is the best for a segment based on the reliability of the post-editor[6] (3 stars: amateur translator, 4 stars: professional translator, 5 stars: certified translator), and on the quality score (from 0 to 20) of MT pre-translations and post-editions.

3.2 Gains in human time (usage value)

From the point of view of the human time spent, how efficient is this method? As the first author is Chinese, he has experimented with French→Chinese and with Chinese→English on segments of a shared TM called Demo2 (including some French and Chinese short articles). We give in Table 2 the statistics gathered during one week (21-27 January 2013). During this week, 1853 segments were post-edited from French into Chinese, and 625 segments were post-edited from Chinese into English, and then an amateur translator translated the same segments without the help of iMAG. We recorded the time taken in each case, and compared the results.

It is well known that one should always post-edit into one's native language: quality should be better and time shorter. However, the measures above seem not to confirm the second point. About 1342/1757=76% of the time is saved in the Fr→Zh direction, and 312/1464=78,6% in the Zh→En direction. But close inspection of the results reveals that, as expected, the Zh post-editions are quite good[7], while the En post-editions are not always exact and very often ungrammatical. Another step of revision by a native English speaker would be necessary before attaining the same translation quality as for Fr-Zh.

Note that these gains are in agreement with early experiments done in 2005 by Jeff Allen[8] with professional translators post-editing into their native language Systran outputs.

We would like to speak here of *usage quality,* or even better of *usage value*. Since the early days where MT was deployed (Hutchins and Somers, 1992), it has been noted that linguistic quality and usage value do not correlate with each other. In fact, *while the linguistic quality of MT outputs is often judged to be very low by linguists and translators, their usage value is often quite high.*

With our setting, the linguistic quality of the post-edited segments (if post-editors work into their native language) is comparable with that of segments translated by junior professional translators having no special knowledge of the "sub-language" of the accessed website[9]. An interesting remark is that, whatever the PE direction, people seem to have some internal sense of "expected speed", that ranges between 15 minutes per page to 25 minutes for the most scrupulous.

[6] 1 star: word for word translation, 2 stars: result of MT.

[7] Five Chinese students to help us verify the results, they proved the correctness and readability.

[8] See http://www.oocities.org/mtpostediting/

[9] The second author has worked as technical translator and revisor and is in a position to make that kind of judgement.

Language pair	Human PE time	Human first draft time	Segments	Source words (Standard pages)	Target words (Standard pages)
French→ Chinese	415 mins	1757 mins	1 853	38 913 (155 p.)	46 648 (116 p.)
Chinese→English	312 mins	1464 mins	625	12 853 (32 p.)	8 568 (34 p.)

Table 2. Statistics from 1-week experiment (we count the number of Chinese characters for Chinese)

Figure 5. Extraction of a "good" TM from a TM produced by "natural" post-edition

Figure 6. Export of a "good" part of a TM

4 Unexpected and costless gains

There are two interesting side effects obtainable without any additional cost: iMAGs can be used to produce high-quality parallel corpora and to set up a permanent task-based evaluation of one or more MT systems

4.1 Production of good quality and "targeted" parallel corpora

Thanks to SECTra_w in-built system of annotation of each translation or post-edition of a segment by a reliability level (from * to *****) and a quality score (0..20), one can extract from the TM associated to a website S a subset verifying any predicate based on levels and scores.

To implement that, we have introduced and implemented into SECTra_w the notion of *selection*. A selection is defined intentionally (by a predicate) or extensionally (by an explicit list), and can be named, for later recall.

Take for example the TM of the website of Greater Grenoble (La Métro) that contains 2500 web pages, or about 30000 segments. More than half have been pre-translated and post-edited into Chinese for the Shanghai Expo in 2010. We may select a "quite good part" of this TM by creating the selection:

```
TM-lametro-extract-good =
TM_select (lametro, [level=3 &
score >=13 | level=4 & score
>=12 | level=5 & score >=11]).
```

The following example shows an even simpler extraction, from the French-Chinese part of the Demo2 TM associated with iMAG-Doc_Par_-jour shown on Figure 5 above. The predicate is

simply [level=3 & score >=13], and its parameters can be directly chosen through the GUI.

The selection obtained can then be exported, as 2 parallel files (source and post-edition) in a simple XML format (Figure 6). SECTra_w also provides additional information (TM, Last updated, Duration of post-editing, post-editor, etc.), and other available download formats (TMX, TXT, and CSV). These data can be used later to "feed" an empirical Moses-based MT system that will become specialized to that website[10].

That possibility is very interesting in the current context. It has been proven that MT systems can be specialized to sublanguages and produce outputs of very high usage value (Chandioux, 1988) (Isabelle, 1987). That means that the outputs are quite readable, and very cheap to post-edit to produce professional quality output.

In recent experiments with a Paris-based multilingual content processing firm, a Moses instance built from a high proportion of a 300K bi-segment TM mixed with a standard parallel corpus extracted from EuroParl (Koehn, 2005) got a BLEU (Papineni et al., 2002) score of about 70%. At this high level, BLEU correlates with usage value: it takes typically 10-15 minutes only to post-edit the equivalent of 1 standard page (250 words, or 400 kanjis), instead of 1 hour to produce a draft translation. But that method works only if a parallel corpus specialized to the sublanguage at hand is available, and that is quite rare in practice.[11]

The situation is similar if the considered MT system is built by an "expert" method (as TAUM-METEO and then METEO).

For example, there is no available parallel Chinese↔French corpus for e-mails, chats, and short technical notes. Building a parallel corpus from scratch is not an option because of the cost of the operation and the scarcity of translators knowing both languages and the technical terms.

Using an iMAG offers a graceful way to solve that difficulty. Whatever MT systems are available, one can begin without any delay to start the bilingual service needed (a web-based chat, for example), routing messages and documents through web pages, and using iMAGs to make them accessible (and improvable) in the desired languages. After a while, the TM-S dedicated to the (empirically defined) sublanguage of S will contain enough "good" bi-segments to extract them and use them to build a specialized instance of an MT system (for example, a specialized Moses-S system[12]).

An important point here is that, in order to encourage end users to post-edit, post-editing should be made very simple and user-friendly. One should refrain from transforming it into a debugging environment for some MT systems. That would also go against the principle to be open to as many MT systems as possible.

5 Continuous task-oriented evaluation of one or more MT systems

The second unexpected benefit of online contributive post-editing using SECTra_w as a backend is that it is possible to directly extract from it objective measures, where references are post-edited MT results.

SECTra_w was initially designed to support an MT evaluation campaign organized by France Telecom R&D (Orange Labs). It includes classical scripts to compute BLEU and NIST (Doddington, 2002), and an original script computing a combination of character-based and word-based edit distances (or semi-distances).

$$\Delta_{comb}(A,B) = c*\Delta_{char}(A,B) + (1-c)\Delta_{word}(A,B) \quad \text{with } 0 \leq c \leq 1.$$

Δ_{char} is computed by the Wagner & Fischer algorithm[13] (Wagner and Fischer, 1974). To compute Δ_{word}, we consider the (typographic) words of strings A and B as a new set of characters, and apply the same algorithm with a matrix M_{word} such that $M_{word}[u, v] = \Delta_{char}[u, v]$. In order *to make the post-editing effort intuitively graspable,* we replace in the W&F Δ_{char} matrix a maximal sequence of N exchanges by N deletions (represented by overstriking and coloring in blue) followed by N insertions (coloring in red).

In evaluation campaigns, one needs to build reference translations, which are produced by expensive professional translators, so that eval-

[10] We are running such an experiment but cannot describe it here for lack of space.

[11] Remember: in 2001, Language Weaver (LW) claimed « to be able to produce an MT system overnight » from a large enough parallel corpus. While that was undoubtedly true, LW produced actually only 4 MT systems in 4 years… because parallel corpora corresponding to the translation needs of solvable clients were and are hard to find.

[12] We have built a French-Chinese Moses system for iMAG-LIG, based on 12000 already post-edited segments.

[13] We use a matrix giving insertion, deletion and exchange costs. Δ_{char} is a distance if all elements are equal to 1, but other values may cause to violate all 3 axioms of distance.

uations are done and redone using the same sets of examples. But, if a website S is post-edited in an access language L, references are produced continuously as contributors (paid or unpaid, organized or occasional) improve the MT pre-translations or the already available post-editions.

Notice also that there is no need whatsoever to PE all segments. PE is normally done by need. If a segment is badly translated but not important or never read, why improve it? That is one aspect of the "multilingual access" concept that makes it intrinsically cheaper than the traditional translation paradigm.

5.1 Evaluating one or more MT systems

Several MT systems can be called on each source segment in L1. When we reconstruct a web page in a target language L2, we choose the best (highest score) and most recent post-edition, if any, or one MT output, for example Systran for En→Zh, or Google Translate for Zh→En.

We can always compute the available similarity measures (or distances, or semi-distances) between the produced post-edition and each of the MT outputs and each of the other post-editions of the segment. The pseudo-trace presented above illustrates that possibility. In this way, each MT system output can be compared against a reference, which is the result of a post-editing activity that is related to the task at hand. In other words, references are produced naturally and with no additional cost.

If (like we do now) the same MT system MT-1 is always chosen as initial value of the string to be modified by post-editing, there is a serious risk of a bias in favour of that system, because of the natural tendency of post-editors to modify the pre-translation as little as possible in order to produce a "good enough" translation (post-edition). Then, whatever the measures used, the outputs of MT-1 will be nearer to the "references" produced by PE than the outputs of the other systems MT-2, MT-3, … , MT-k.

How to improve on that? A first idea would be to ask k humans to post-edit all k MT outputs. But that would multiply by k the human time taken, and it would clearly be quite unrealistic if one wants to integrate evaluation in a task-related activity without additional cost.

In the future, we plan to choose (automatically) among the k possible MT outputs so that each MT-k is guaranteed to be used for a fixed proportion of the segments. The simplest way is to "rotate" between systems (choice (n) = n modulo k), so that n/k of inputs will be pre-translated by each MT system. It is also possible to "throw the dice", so that each of MT-1, … , MT-k will have 100/k % chances to be chosen. There may also be good reasons to give more chances to one MT system, for example to a system being developed and still at the beginning of its "learning curve". The rotation and controlled random choice methods above can easily be adapted to that idea.

6 Conclusion and perspectives

In this paper we have shown that an interactive Multilingual Access Gateway (iMAG) dedicated to a website S (iMAG-S) is quite helpful to make S accessible in many languages immediately and without editorial responsibility. Visitors of S contribute to the continuous and incremental improvement of the most important textual segments, and eventually of all. In this approach, pre-translations are produced by one or more MT systems. To have all (100%) segments post-edited is *not* the goal: it is quite OK if post-edited segments are only those that are important (often accessed) and badly MT-translated.

Continuous use since 2008 on many websites and for several access languages shows that a quality comparable to that of a first draft by junior professional translators is obtained in about 40% of the (human) time, sometimes less, with the condition that contributors post-edit into their native language.

An interesting observation is that post-editors seem to have some kind of personal "expected PE speed" that does not depend on the direction of post-editing. The resulting quality, then, depends only on their expertise in each direction. Note that, although in principle counter-indicated, post-editing from one's mother tongue may be cost-effective for some situations like Chinese-French in a French firm: acceptable quality at a still reasonable cost can be obtained by PE first the result of Chinese-English MT by a Chinese, and then the result of English-French MT by a French.

We have also shown and illustrated two interesting side effects obtainable without any added cost: an iMAG-S can be used to produce a high-quality parallel corpus and to set up a permanent task-based evaluation of one or more MT systems. By nature, the HQ parallel corpus extractable from a TM-S is specialized to the sub-

language of the website S. When it becomes large enough after some period of using the iMAG-S (about 10-15000 'good' bi-segments for the sublanguages of classical web sites), it can be used to build an empirical MT system for that sublanguage, and then to improve it incrementally as time goes and new segments are post-edited. Recent experiments in specializing empirical MT systems have shown that remarkably good MT results can be obtained (Rubino et al., 2012). We are running an experiment on French-Chinese that seems to confirm it.

Acknowledgments

This work has been partially funded by UJF, ANR (Traouiero project), L&M (Lingua et Machina) and ANRT.

References

C-P. Huynh, C. Boitet, and H. Blanchon. 2008. SECTra_w: an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proc. LREC-08, demonstration session*, 8 p., Marrakech, 27-31/5/08, ELRA/ELDA, ed.

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proc. HLT 2002*. vol. 1/1: pp. 128-132 (notebook proceedings). San Diego, California. March 24-27, 2002.

H-P. Nguyen, C. Boitet, and G. Sérasset. 2007. PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot. *SNLP- 2007,* 6 p, Bangkok, Thailand, 2007.

J. Chandioux. 1988. 10 ans de METEO. Traduction Assistée par ordinateur. *Actes du séminaire international sur la TAO et dossiers complémentaires*, OFIL, A. Abbou, ed. Paris.

K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *In 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, PA.

M. Daoud, C. Boitet, A. Kitamoto, and M. Mangeot. 2009. Building a Community-Dedicated Preterminological Multilingual Graphs from Implicit and Explicit User Interactions. *Second International Workshop on REsource Discovery (RED 2009)*, co-located with VLDB 2009, Lyon, France, 8 p.

P. Isabelle. 1987. Machine translation at the TAUM group. Machine Translation: *The State of the Art*, pp. 247–277.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *In Proceedings of the Annual Meeting of the ACL, demonstration session,* pp. 177–180, Prague, Czech Republic.

P. Koehn. 2005. Europarl: a Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: the tenth Machine Translation Summit,* pp. 79-86, AAMT, Phuket, Thailand, 2005.

R-A. Wagner, and M-J. Fischer. 1974. The String-to-String Correction Problem. *JACM 21*: pp. 168-173.

R. Rubino, S. Huet, F. Lefèvre, and G. Linarès. 2012. Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique, *In Conférence en Traitement Automatique des Langues Naturelles*, pp. 527-534, Grenoble, France.

W-J. Hutchins, and H-L. Somers. 1992. An introduction to machine translation. *Academic Press*, pp. 175-177, London.