User-focused task-oriented MT evaluation for wikis: a case study

Federico Gaspari, Antonio Toral and Sudip Kumar Naskar
School of Computing
Dublin City University
Dublin 9, Ireland
{fgaspari, atoral, snaskar}@computing.dcu.ie

Abstract

This paper reports on an evaluation experiment focusing on statistical machine translation (MT) software integrated into a more complex system for the synchronization of multilingual information contained in wiki sites. The experiment focused on the translation of wiki entries from German and Dutch into English carried out by ten media professionals, editors, journalists and translators working at two major media organizations who post-edited the MT output. The investigation concerned in particular the adequacy of MT to support the translation of wiki pages, and the results include both its success rate (i.e. MT effectiveness) and the associated confidence of the users (i.e. their satisfaction). Special emphasis is laid on the post-editing effort required to bring the output to publishable standard. The results show that overall the users were satisfied with the system and regarded it as a potentially useful tool to support their work; in particular, they found that the post-editing effort required to attain translated wiki entries in English of publishable quality was lower than translating from scratch.

1 Introduction

1.1 Background to the evaluation

The evaluation reported in this paper was conducted as part of the three-year CoSyne project, funded by the EU under the FP7 scheme.[1] The project consortium includes seven partners: three academic institutions, i.e. University of Amsterdam (UvA, The Netherlands) as coordinator, Fondazione Bruno Kessler (FBK, Italy) and Dublin City University (DCU, Ireland); one research organization, the Heidelberg Institute for Theoretical Studies (HITS, Germany); and three end-user partners, Deutsche Welle (DW, Germany), the Netherlands Institute for Sound and Vision (NISV, The Netherlands) and Wikimedia Foundation Netherlands (WMF).

The aim of the technology developed within the CoSyne project is to facilitate the synchronization of the contents of wiki sites across different languages. In this context, machine translation (MT) is used as part of an integrated system which includes other modules that take care of textual entailment, document structure modeling, overlap synchronization (to check how similar or different wiki entries on the same topic in multiple languages are), insertion point detection (to establish where new machine-translated information should be added or replaced following some edits), etc.

The statistical MT software incorporated within the CoSyne system was developed by UvA (Martzoukos and Monz, 2010). Toral et al. (2011) conducted a comparative evaluation of the CoSyne MT software against four free web-based MT systems over data sets from the news domain based on a range of state-of-the-art automatic metrics. In its first year up to February 2011, the project has covered three language pairs, i.e. German, Dutch and Italian from and into English. By the end of the project, Turkish and Bulgarian will also be included to extend

[1] More details of the project are available at www.-cosyne.eu

the relevance of the CoSyne system to languages for which fewer resources are available. This paper focuses in particular on the evaluation of the CoSyne MT system to translate wiki entries in the German→English and Dutch→English language directions in real-life scenarios, based on the feedback of actual users.

1.2 Structure of the paper

The rest of the paper has the following structure. Section 2 provides a brief overview of the state-of-the-art in MT evaluation, considering in particular the user-focused perspective. Section 3 explains the scenario in which the experiment reported here was conducted, while Section 4 describes the format and contents of the questionnaire that was administered to the users for the evaluation. Section 5 presents the results and discusses the most interesting findings. Finally, Section 6 draws some conclusions and outlines plans for future work.

2 Related work

Evaluation has been a central concern in the field of MT virtually since its beginnings (White, 2003). Judging the correctness of the translation might intuitively be considered the obvious method to evaluate MT output. However, this has proved to be too broad to define unambiguously, and it is quite common to break the concept of translation correctness down into two sub-criteria, i.e. fluency (does the output read well?) and adequacy (does the output preserve the meaning of the input?) (Koehn, 2009). A substantial effort to standardize MT evaluation was carried out as part of the ISLE project, which resulted in FEMTI, a framework that defines the possible evaluation requirements and system characteristics to be evaluated (King et al., 2003).

Nowadays, MT research and development, especially within the statistical paradigm, are crucially dependent on fast and cheap automatic MT evaluation metrics. Research in MT evaluation has gained momentum in the last decade, and as a result a number of such automatic MT evaluation metrics are widely used today, e.g. BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) to mention two of the most popular.

However, automatic evaluation metrics are relatively crude and do not provide any insight into the nature and severity of the errors. This means that they do not distinguish between different types of errors that have various implications depending on the task at hand, particularly in terms of the post-editing effort that they entail if high quality is required for the final translation. Moreover, the correlation of automatic evaluation metrics with human judgment is not optimal and, interestingly, high correlation in both adequacy and fluency cannot always be achieved at the same time (Lin and Och, 2004).

As a consequence, human evaluation arguably still provides the most reliable method to judge MT output, despite being tedious and expensive. As a matter of fact, questionnaire-based manual evaluations of the two aforementioned criteria, fluency and adequacy, have been carried out at WMT, both scoring sentences of an MT system using a graded scale and ranking the quality of the output provided by two or more systems (Koehn and Monz, 2005). Following a well-established tradition in MT evaluation, the approach taken for this study focused primarily on the utility of the MT output and on the users' level of satisfaction (White and O'Connell, 1996).

In connection with this, we also considered post-editing effort and usefulness (Krings, 2001; Allen, 2003; O'Brien, 2007), particularly to gain an insight into whether the deployment of the CoSyne system was regarded as useful compared to having to translate the wiki entries from scratch. There is a body of work (Guerra Martínez, 2003; O'Brien, 2005; Gueberof, 2009; Koehn and Haddow, 2009; Specia and Farzindar, 2010; Carl et al., 2011; Specia, 2011) that has recently studied this particular aspect, i.e. whether post-editing MT output is considered easier than translating from scratch.

3 User-focused task-oriented MT evaluation for wikis

3.1 Overall evaluation framework

The evaluation exercise that forms the basis of this paper was conducted before the mid-way mark of the three-year project, with a full-scale final evaluation envisaged at the very end of the project. Hence, the initial round of end-user evaluation described here can be seen as a pilot study. During this process we have learned valuable lessons concerning the complexities of evaluating MT software which is part of a wider system that is used to optimize translation in highly dynamic wiki environments with the support of post-editing.

The ongoing evaluation effort undertaken as part of the project also covers other aspects of the overall CoSyne system, including textual entailment, document structure modeling and induction, as well as usability and interaction design. However, for the purposes of this paper we restrict the data analysis and the discussion of the results exclusively to the MT dimension, looking at the success of the MT component and at the satisfaction on the part of its users, who post-edited the raw MT output to obtain translated wiki pages of publishable quality in English.

3.2 Evaluation scenario

A protocol for the evaluation exercise was agreed upon between DCU, as the technical partner responsible for this area of the project, and the two end-users planning to deploy the CoSyne system in the future, that is to say DW and NISV, who provided the users from their staff. The evaluation sessions took place at their respective premises, and were run by local senior staff who are also CoSyne project members. To ensure similar experimental set-ups and conditions so as to safeguard the comparability of the results, the protocol stipulated a number of requirements, the most important of which are described in what follows. Therefore, the results of the experiment for the two language pairs covered in the evaluation between the two groups of users are broadly comparable, and will be analyzed in Section 5.

DW and NISV selected staff with German-English and Dutch-English as their working languages, respectively, among their editors, journalists, translators and project managers to participate in the evaluation. These are the groups of professionals who are envisaged to take advantage of the final CoSyne system for the multilingual synchronization of wiki content, once it is deployed. Similarly, the evaluation exercise was conducted on wiki entries reflecting the typical texts that the two organizations need to translate into English.

The source chosen by DW for this was "Today in History",[2] a bilingual German/English wiki-style website which has entries referring to salient historical events organized by date (including the dates of birth of important or famous people — the more extensive German section is called "Kalenderblatt"), while the environment chosen by NISV was their own wiki site containing media-related information primarily of interest to the Dutch public (mostly concerning TV and cinema, including profiles of programmes and series, biographies of actors and directors, etc., and currently available only in Dutch).[3]

It was agreed that the evaluation should take place by means of a written questionnaire in English (see Section 4 for more details on its structure and contents). The users involved were asked to focus their responses exclusively on the linguistic quality and the level of usefulness of the MT tool as such, in isolation from the other components of the CoSyne system, i.e. regardless of the performance of textual entailment, document structural analysis, usability and interaction design of the interface, etc. (we recognize, however, that this involves some degree of approximation, insofar as these issues may have affected the success of the translation and post-editing processes, as judged by the users).

A time-tracking system was implemented, to gather data regarding the time spent by each user post-editing each wiki entry. The users were made aware of this, and asked to complete the sessions without taking breaks during the post-editing of individual wiki entries. In addition, the post-editing changes performed by the participants to improve the raw MT output in English provided by the CoSyne MT software were logged.

Before starting the evaluation, the participants were given a short presentation of the CoSyne project and of the system that is being developed within its framework, including a brief demo of the main functionalities of the first prototype. They were then asked to experiment with the CoSyne system for a period of one to three hours (depending on the typical

[2] "Today in History/Kalenderblatt" is available at `http://www.todayinhistory.de`

[3] The NISV wiki is available at `http://www.beeldengeluidwiki.nl`

length of the wiki entries to be translated at the two end-user partners). This was done to ensure that they filled in the questionnaires after having similar experience with typical use cases, before judging the MT quality and commenting on the post-editing effort.

4 The instrument: structure of the evaluation questionnaire

This section describes the questionnaire[4] that was used for the evaluation of the CoSyne MT software. The questions were all formulated in English, given that this was the common language for all the respondents, and grouped into parts focusing on different aspects of the evaluation. The questionnaire included approximately 50 items using different formats, such as Likert scale, multiple choice and open questions.

Part A of the questionnaire covered basic demographic information about the respondents, such as age, gender, language background, role held in the organization and level of seniority, etc. Part B concerned the previous use of MT on the part of the users, to clarify with what background experiences of translation technology they approached the evaluation exercise. If the participants had already used MT in the past, they were asked further questions to elucidate which systems they had previously used and for what purposes. Part C of the questionnaire focused specifically on the users' impressions of the CoSyne MT software during the evaluation session. This was followed by part D, regarding the post-editing work undertaken by the participants to bring the raw MT output in English of the wiki entries to publishable standard. Finally, part E elicited some general comments and impressions from the users about their experience with the CoSyne MT software. In addition, six questions in a final part focused on the usability and interaction design of the overall CoSyne system. Since these answers were not directly related to MT evaluation for wiki content, they are not covered in this paper.

5 Results and discussion

5.1 Demographic information about the participants (part A)

Out of the ten users who were involved in this run of the evaluation, six were DW staff, and four work at NISV. Between these two groups, there was a slight gender imbalance, with six men and four women taking part (question A4). Their average age (A3) was 34, with the youngest respondent being 20, and the oldest 46. They covered a variety of roles (A5) involving content generation for the wiki sites in English at DW or NISV, i.e. editors, authors, translators and project managers. On average they had been working at their organization for just over 3 years (A6), but with different levels of seniority, which for DW ranged from a recent intern to a freelance author with ten-year experience.

All four NISV staff were native speakers of Dutch, while the DW users included one native speaker of Romanian fluent in German, with all the others being native speakers of this language (A7). 80% of the participants self-rated their knowledge of English as upper-intermediate, with 20% defining it as either intermediate or excellent (A8). Interestingly, none of the users considered themselves to be bilingual.

5.2 Previous use of MT (part B)

This part of the questionnaire focused on past experiences that the users might have had with any MT system prior to taking part in the evaluation, which was useful to establish their expectations based on previous encounters with MT. 80% of the participants said that they had already used MT at some point in the past before the experiment (B1). In particular, 7 of them had used it for personal reasons, and 6 had also employed MT software as part of their work (B2). All but one of them had used Google Translate, while the remaining respondent had only used Babel Fish, and two participants reported using both these free online MT services (B3).

4 respondents had used MT for translating from English into other languages (for example Vietnamese), and 6 for translating into English from a range of source languages. 5 of them reported experiences with other language combi-

[4] The questionnaire is available at `http://www.computing.dcu.ie/~atoral/cosyne/quest.pdf`

nations not involving English, such as from Italian to Dutch, German to Swedish, German–Spanish in both directions, and finally Arabic, Chinese and Serbian to German (B4). In all these cases, the use of MT for assimilation purposes was quite common (75%), with a lower usage for dissemination purposes (25%). Interestingly, 62.5% of the respondents stated that they had post-edited MT output to obtain high-quality translations from the initial raw draft provided by the system that they had used (B5).

With regard to the kind of materials that they had translated by means of MT (B6), 2 users mentioned business correspondence and professional or personal emails. Other responses concerned online articles (3 individuals) and websites in general (2 cases, with one user remarking that "the translations of Dutch sites to English were hilarious"); one of these 2 users specified that they had used online MT in the past to translate Wikipedia content. Finally, 3 respondents stated that they had previously resorted to web-based MT to assist their studies or to read academic and university-related papers, and a couple of others mentioned contracts and technical documents when describing the kinds of texts for which they had deployed MT.

Based on these mixed experiences, overall the 8 respondents who provided answers in this respect had a predominantly negative-to-neutral impression of MT quality before taking part in the evaluation of the CoSyne MT system, based on a 5-point Likert scale (results are shown in Table 1 – with scores ranging from 1 "very poor" to 5 "very good"). In particular, 1 of them (12.5%) described their impression of the MT quality they had experienced in the past as "very poor", 2 (25%) as "poor", 3 (37.5%) as "neither poor nor good", and finally, 2 respondents (25%) regarded it as "good", with none giving "very good" as an answer (B7). The average (2.8) is just below the medium value (3). The average for German→English is clearly higher than that for Dutch→English (3.2 versus 2) with the same deviation (0.8). Better scores for German→English were found consistently in the answers to all the remaining questions.

These initial modest impressions can be usefully compared to how all the ten respondents evaluated the quality of the CoSyne MT software to translate wiki pages, looking at the results discussed in Section 5.3.

	Quality
Very poor (1)	12.5%
Poor (2)	25.0%
Medium (3)	37.5%
Good (4)	25.0%
Very good (5)	0.0%
Average	2.8
de→en	3.2
nl→en	2.0
Deviation	1.0
de→en	0.8
nl→en	0.8

Table 1: Quality of previously used MT systems

5.3 Use of the CoSyne MT software (part C)

Following on from their previous experience with other MT systems, the respondents were then asked to evaluate the performance of the CoSyne MT component, that they had applied to the translation of wiki entries into English. Table 2 shows the quality (C2) and usefulness (C3) of the CoSyne MT component as it was perceived by the users. The average quality is considered exactly medium (3), which interestingly is better than the perceived quality for previously used systems (B7, 2.8). The usefulness is slightly higher than medium (3.3).

	Quality	Usefulness
Very poor (1)	10%	0%
Poor (2)	20%	30%
Medium (3)	30%	20%
Good (4)	40%	40%
Very good (5)	0%	10%
Average	3.0	3.3
de→en	3.7	3.8
nl→en	2.0	2.5
Deviation	1.1	1.1
de→en	0.5	1.0
nl→en	0.7	0.5

Table 2: Quality and usefulness of the CoSyne MT system

Table 3 (C4) indicates whether using the MT system helps in obtaining the translation more quickly than translating manually from scratch. The average value (4.6) of the answers is slightly higher than the mid-point of the scale (4), thus favoring the MT system, in line with the findings presented by Plitt and Masselot (2010) and Flournoy and Rueppel (2010). It should also be noted that there is a clear spread of answers, as corroborated by the high value of the deviation (2.1).

Strongly disagree (1)	20%
	0%
	0%
	10%
	30%
	30%
Strongly agree (7)	10%
Average	4.6
de→en	5.7
nl→en	3.0
Deviation	2.1
de→en	0.8
nl→en	2.1

Table 3: Is MT faster than translation from scratch?

The users identified the linguistic phenomena that caused most problems in translation, both in the source and target languages (C5 and C6, respectively). Some of these are common to both source and target: proper nouns, syntax, pronouns and verbs, the main problem for the latter being that they are frequently dropped in the translation. Problems due to the source language are different for German and Dutch. For the former, mistranslations occur most commonly for compounds, and there are problems with subordinate clauses, word order and unknown words. For Dutch, problems typically arise with prepositions, number agreement, and the rendition of idioms and figures of speech. Regarding the target language, the main problems are due to word order, unknown words and capitalization.

Table 4 presents the results regarding MT quality (C7) according to five aspects (**accu**racy, **corr**ectness, **comp**rehensibility, **read**ability and **styl**istic appropriateness). Only accuracy (3.6) is above average (3.5). Comprehensibility and readability lag a bit behind (3.2), while correctness (2.7) and style (2.6) obtain lower scores. It should be noted, however, that none of the average values for the evaluation of these five criteria came up as particularly poor, considering both language pairs together. For question C7 we decided to break down the global notion of quality into these related features to be as fine-grained as possible in our investigation. We recognize, however, that some of these parameters overlap, and we did not explain the detailed differences between them to the respondents, who might not have appreciated the subtle differences involved.

	accu	corr	comp	read	styl
Poor (1)	10%	10%	10%	10%	0%
	0%	30%	20%	20%	40%
	50%	40%	30%	40%	60%
	10%	20%	30%	10%	0%
	20%	0%	0%	10%	0%
	10%	0%	10%	10%	0%
Excellent (7)	0%	0%	0%	0%	0%
Average	3.6	2.7	3.2	3.2	2.6
de→en	4.2	3.3	3.8	4.0	2.8
nl→en	2.8	1.8	2.3	2.0	2.3
Deviation	1.4	0.9	1.4	1.5	0.5
de→en	1.3	0.5	1.2	1.3	0.4
nl→en	1.1	0.4	1.1	0.7	0.4

Table 4: MT quality breakdown

5.4 Post-editing the MT output for wiki translation (part D)

Table 5 shows the amount of work required to post-edit the output of the MT system in terms of time (D1) and effort (D2), as perceived by the users themselves. For both languages taken together, the average values for time and effort (4.7 and 4.5, respectively) are higher than the medium value (3.5), which shows that the users thought that the raw MT output requires substantial work to be made publishable. It should be noted, however, that post-editing translations from Dutch was considered to impose a much higher workload than the other language pair.

Table 6 shows how often the users felt that they needed to refer to the source language in order to post-edit the translation (D3). This result (averaging 5.8) shows that the users very often consulted the source text in order to post-edit the raw MT output, and the NISV users tended to do this more often than their DW counterparts.

	Time	Effort
Short/small (1)	0%	0%
	0%	20%
	10%	0%
	40%	10%
	20%	50%
	30%	20%
Long/large (7)	0%	0%
Average	4.7	4.5
de→en	4.3	3.8
nl→en	5.3	5.5
Deviation	1.1	1.4
de→en	1.0	1.5
nl→en	0.8	0.5

Table 5: Work required to post-edit

Never (1)	0%
	10%
	0%
	0%
	10%
	50%
Always (7)	30%
Average	5.8
de→en	5.3
nl→en	6.5
Deviation	1.5
de→en	1.8
nl→en	0.5

Table 6: Need to refer to the source language

Tables 7 and 8 show the judgments in terms of severity and frequency, respectively (D4 to D7), of the different post-editing operations (**ins**ertions, **del**etions, **sub**stitutions and **reo**rderings). Looking at all the ten participants together, severity of deletions (3.8) is considerably lower than the other operations, which obtain similar values in the range [4.5-5]. The perceived frequency is very similar for all the operations, with values ranging between 4.7 and 5.1.

	ins	del	sub	reo
Irrelevant (1)	0%	0%	0%	0%
	20%	40%	10%	20%
	0%	0%	10%	0%
	0%	30%	30%	0%
	50%	10%	30%	30%
	20%	10%	10%	40%
Very serious (7)	10%	10%	10%	10%
Average	4.8	3.8	4.5	5.0
de→en	4.2	2.8	3.8	4.2
nl→en	5.8	5.3	5.5	6.3
Deviation	1.6	1.8	1.4	1.7
de→en	1.7	1.3	1.2	1.7
nl→en	0.8	1.3	1.1	0.4

Table 7: Severity of errors over post-editing operations

	ins	del	sub	reo
Absent (1)	0%	0%	0%	0%
	10%	10%	0%	10%
	10%	0%	0%	0%
	10%	40%	30%	10%
	40%	30%	50%	40%
	10%	0%	0%	30%
Frequent (7)	20%	20%	20%	10%
Average	4.9	4.7	5.1	5.1
de→en	4.3	4.0	4.5	4.3
nl→en	5.8	5.8	6.0	6.3
Deviation	1.6	1.5	1.1	1.4
de→en	1.5	1.1	0.5	1.2
nl→en	1.3	1.3	1.0	0.4

Table 8: Frequency of errors over post-editing operations

5.5 Final comments (part E)

This part asked the users about the most important positive aspects (E1), weaknesses (E2) and any other comments (E3) that they wanted to point out at the end of the evaluation of the CoSyne MT software. The positive aspect that was mentioned most frequently was the provision by the system of a draft translation to work upon. Other valuable aspects mentioned were the integration in the wiki environment and the potential to speed up the translation task.

The main weakness regarded the translation quality, mainly concerning the wrong translation of pronouns and verbs that were frequently dropped, incorrect word order, mistranslated compounds and limited lexical coverage. Most of the general comments praised the potential of the system, acknowledging the achievements shown by the first prototype.

6 Conclusions and future work

This paper has presented a user-focused task-oriented evaluation framework for MT by means of a questionnaire specifically designed for this aim. This has been applied to evaluate the first prototype of the MT component of the CoSyne system, used to synchronize entries in multilingual wikis for the German→English and Dutch→English language directions.

One finding worth stressing is that the quality of the MT system evaluated, as perceived by the users, is higher than the quality perceived for previously used MT systems. This might indicate that the current system produces better translations, but it is also plausible that the users' lower judgment of previously used MT systems is due to a negative bias towards this technology, which might be mitigated when they actually use it.

On another note, although the effort required to post-edit raw MT output in order to achieve publishable content is deemed to be high, the users still found that it takes less time to translate text by post-editing MT output compared to translating it manually from scratch. This was especially the case for the DW staff working on the German into English language direction.

The breakdown of the results for the two language pairs considered consistently shows that translations from German are rated more favorably than those from Dutch. As a matter of fact, the answers focusing on errors regarding post-editing operations indicated a worse performance for Dutch→English by around 20% in a 7-point scale. This contrasts with earlier findings in an evaluation of the CoSyne MT system using eight state-of-the-art automatic metrics (Toral et al., 2011), where Dutch→English scored higher than German→English for all these metrics. This warrants further investigation into this discrepancy – e.g. we cannot rule out that the two groups of users approached the evaluation tasks with different expectations, nor that some variation in these results can be attributed to differences in the lexical and stylistic properties of the texts found in the wiki sites in the two languages.

In terms of future work building on this initial study, we plan to extend the analysis presented here by looking into the logs of the actual edits performed by the users on the wiki entries after MT processing, and considering the amount of time that the professionals spent post-editing (this information has been recorded by the end-user partners), to estimate the costs of post-editing MT output for the multilingual dissemination of synchronized wiki contents.

We are also planning to study more closely the correlation between the post-editing carried out by the users on the one hand, and the results provided by the automatic evaluation metrics TER and TERp on the other. These two metrics indicate insertions, deletions, substitutions and shifts required to match the reference translation (Snover et al., 2006; Snover et al., 2009), so it will be interesting to investigate the extent to which their results correspond to the actual post-editing work carried out by the MT users.

Finally, as part of the CoSyne project, the DCU team has developed DELiC4MT[5] (Naskar et al., 2011), an automatic diagnostic evaluation tool which detects classes of linguistic phenomena in the input (source-language) text giving rise to errors in the MT output, compared to human reference translation. As part of our future work we intend to use DELiC4MT to monitor the performance of the MT system on the linguistic phenomena flagged up as most problematic by the users during post-editing.

Acknowledgments

This work has been funded in part by the European Commission through the CoSyne project FP7-ICT-4-248531. The authors are grateful to the staff at Deutsche Welle and at the Netherlands Institute for Sound and Vision who took part in the evaluation sessions, and particularly to Peggy van der Kreeft and Tim Koch (DW) and Jaap Blom (NISV) who organized and supervised these sessions in which the data used for this study were collected; in addition, Christophe Costa Florêncio and Christof Monz (University of Amsterdam) commented on an

[5] DELiC4MT is available at `http://www.computing.dcu.ie/~atoral/delic4mt` (under the GPL-v3 license).

earlier draft of this paper. Any remaining errors are the sole responsibility of the authors.

References

Allen, Jeff (2003) "Post-editing". Somers, Harold (ed). *Computers and Translation: a translator's guide*. Amsterdam: John Benjamins, pp. 297-317.

Banerjee, Satanjeev and Alon Lavie (2005) "An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". *Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan, pp. 65-72.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt Lykke Jakobsen (2011) "The process of post-editing: a pilot study". Sharp, Bernadette, Michael Zock, Michael Carl and Arnt Lykke Jakobsen (eds). *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation.* Copenhagen Business School, Copenhagen, Denmark, 20-21 August 2011. Frederiksberg: Samfundslitteratur, pp.131-142.

Flournoy, Raymond and Jeff Rueppel (2010) "One Technology: Many Solutions". *Proceedings of AMTA 2010: the Ninth conference of the Association for Machine Translation in the Americas*. Denver, Colorado, USA, October 31-November 4, 2010, 6 pages.

Gueberof, Ana (2009) "Productivity and quality in MT post-editing". *Proceedings of MT Summit XII Workshop: Beyond Translation Memories. New Tools for Translators.* Ottawa, Ontario, Canada, 29 August 2009, 9 pages.

Guerra Martínez, Lorena (2003) *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output.* Unpublished MA Dissertation, Dublin City University, Dublin, Ireland.

King, Margaret, Andrei Popescu-Belis and Eduard Hovy (2003) "FEMTI: creating and using a framework for MT evaluation". *Proceedings of the IX Machine Translation Summit*. New Orleans, Louisiana, USA, pp. 224-231.

Koehn, Philipp and Christof Monz (2005) "Shared task: statistical machine translation between European languages". *Proceedings of the ACL Workshop on Building and Using Parallel Texts (ParaText '05)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 119-124.

Koehn, Philipp (2009) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, Philipp and Barry Haddow (2009) "Interactive assistance to human translators using statistical machine translation methods". *Proceedings of MT Summit XII.* Ottawa, Ontario, Canada, 26-30 August 2009, pp. 73-80.

Krings, Hans P. (2001) *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Geoffrey S. Koby (ed). Kent, OH: Kent State University Press.

Lin, Chin-Yew and Franz Josef Och (2004) "ORANGE: a method for evaluating automatic evaluation metrics for machine translation". *Proceedings of the 20th international conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 501-507.

Martzoukos, Spiros and Christof Monz (2010) "The UvA system description for IWSLT 2010". *Proceedings of the 7th International Workshop on Spoken Language Translation*. Paris, France, pp. 205-208.

Naskar, Sudip Kumar, Antonio Toral, Federico Gaspari and Andy Way (2011) "A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints". To appear in *Proceedings of Machine Translation Summit XIII*. Xiamen, China, 19-23 September 2011, pp. 529-536.

O'Brien, Sharon (2005) "Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability". *Machine Translation* 19(1), pp. 37-58.

O'Brien, Sharon (2007) "An empirical investigation of temporal and technical post-editing effort". *Translation and Interpreting Studies* 2(1), pp. 83-136.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (2002) "BLEU: a method for automatic evaluation of machine translation". Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, PA, pp. 311-318.

Plitt, Mirko and François Masselot (2010) "A productivity test of statistical machine translation post-editing in a typical localization context". *Fourth Machine Translation Marathon "Open Source Tools for Machine Translation", 25-30 January, Dublin, Ireland. Prague Bulletin of Mathematical Linguistics* 93 (January 2010), pp. 7-16.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul (2006) "A Study of Translation Edit Rate with Targeted Human Annotation". *Proceedings of the 7th Confer-*

ence of the Association for Machine Translation in the Americas (AMTA-2006), “Visions for the Future of Machine Translation”. Cambridge, Massachusetts, USA, 8-12 August 2006, pp. 223-231.

Snover, Matthew, Nitin Madnani, Bonnie Dorr and Richard Schwartz (2009) “Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric”. *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009).* Athens, Greece, 30-31 March 2009, pp. 259-268.

Specia, Lucia and Atefeh Farzindar (2010) “Estimating machine translation post-editing effort with HTER”. *Proceedings of JEC 2010: Second joint EM+/CNGL Workshop “Bringing MT to the user: research on integrating MT in the translation industry”, AMTA 2010.* Denver, Colorado, November 4, 2010, pp. 33-41.

Specia, Lucia (2011) “Exploiting objective annotations for measuring translation post-editing effort”. Forcada, Mikel L., Heidi Depraetere and Vincent Vandeghinste (eds). *Proceedings of the 15th International Conference of the European Association for Machine Translation*. Katholieke Universiteit Leuven, Leuven, Belgium, 30-31 May 2011, pp. 73-80.

Toral, Antonio, Federico Gaspari, Sudip Kumar Naskar and Andy Way (2011) “A Comparative Evaluation of Research vs. Online MT Systems”. Forcada, Mikel L., Heidi Depraetere and Vincent Vandeghinste (eds). *Proceedings of the 15th International Conference of the European Association for Machine Translation*. Katholieke Universiteit Leuven, Leuven, Belgium, 30-31 May 2011, pp. 13-20.

White, John S. and Theresa A. O'Connell (1996) “Adaptation of the DARPA Machine Translation Evaluation Paradigm to End-to-End Systems”. *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Quebec, pp. 106-114.

White, John S. (2003) “How to evaluate machine translation”. Somers, Harold (ed). Computers and Translation: a translator's guide. Amsterdam: John Benjamins, pp. 211-244.