Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles

Sheila C. M. de Sousa, Wilker Aziz and **Lucia Specia**
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
{sheila.castilhomonteirodesousa, w.aziz, l.specia}@wlv.ac.uk

Abstract

With the increasing demand for fast and accurate audiovisual translation, subtitlers are starting to consider the use of translation technologies to support their work. An important issue that arises from the use of such technologies is measuring how much effort needs to be put in by the subtitler in post-editing (semi-)automatic translations. In this paper we present an objective way of measuring post-editing effort in terms of *time*. In experiments with English-Portuguese subtitles, we measure the post-editing effort of texts translated using machine translation and translation memory systems. We also contrast this effort against that of translating the texts without any tools. Results show that post-editing is on average 40% faster than translating subtitles from scratch. With our best system, more than 69% of the translations require little or no post-editing.

1 Introduction

Automatic and semi-automatic translation have become a potential help in the subtitling industry due to the increasing demand for translations and the short time professionals have to deliver them. Many attempts have been made to translate subtitles automatically by using different Machine Translation (MT) approaches such as Rule-Based (RBMT), Example-Based (EBMT), Statistical (SMT) and also Translation Memory (TM) systems. However, no previous work compares different approaches in terms of the effort that is required to post-edit the translations they produce. Additionally, the related work in the field does not provide an in-depth comparison between the effort needed to translate subtitles from scratch and the effort needed to post-edit a draft version produced using translation tools.

The ability to objectively assess translation technology tools according to their post-editing effort is essential for a well informed decision among the large variety of tools available, as well as to ensure that such tools produce translations that require less effort to post-edit (PE) than the effort that would be necessary to translate the same texts from scratch (HT).

In this paper we compile a corpus of English – Brazilian Portuguese subtitles and we compare two different MT approaches as well as a TM system using this corpus. The translations obtained are post-edited and the original sentences are also translated from scratch, both using a tool specially designed to gather objective and subjective effort indicators: time spent on performing the task and qualitative assessments. Results show that translators can greatly benefit from automatically obtained translations.

The rest of this paper is organized as follows: Section 2 gives an overview of prior work; Section 3 describes our parallel corpus of subtitles; Section 4 describes how the experiments were performed; Section 5 presents the results; and Section 6 concludes the paper and gives some directions for further research.

2 Related Work

Popowich et al. (2000) propose a number of preprocessing steps in order to improve the accuracy of an RBMT system for translating closed captions. Two native speakers assessed the translations, reporting 70% accuracy.

O'Hagan (2003) experiments with English-Japanese subtitles for the movie *The Lord of the Rings*. Subtitles from the first movie are used to feed a TM system and subtitles from the second movie are used for testing. Results are not encouraging, probably due to the poor TM coverage.

Proceedings of Recent Advances in Natural Language Processing, pages 97–103,
Hissar, Bulgaria, 12-14 September 2011.

Armstrong et al. (2006) train an EBMT system in two scenarios: i) using a homogenous corpus compiled exclusively with DVD subtitles, and ii) using a heterogenous corpus compiled with a mix of subtitles and sentences from the Europarl (Koehn, 2005). The results show that a homogenous setting leads to better translations.

Flanagan (2009) extends the work of Armstrong et al. (2006) by using larger parallel corpora of subtitles from multiple genres. A subjective evaluation querying users who watched movies containing the translated subtitles in terms of intelligibility and acceptability was performed. Results show an average performance ($\sim$ 3 on a 1-6 scale).

Melero et al. (2006) combine a black-box MT system and a TM using a corpus of newspaper articles and United Nation texts to translate subtitles. They find that MT+TM performs significantly better than MT in terms of BLEU (Papineni et al., 2002) in an English-Spanish task. For English-Czech they compare HT against PE in terms of time. The comparison is somewhat inconclusive as the HT and PE were compared using different texts and a single human translator.

Volk (2008) uses a large proprietary corpus of subtitles (5 million sentences) to train an SMT system. The author reports BLEU: i) using a single reference, and ii) using the translations produced by six post-editors. The author finds that SMT outputs can still be acceptable translations even though they do not exactly match the HT as long as they lie within 5 keystrokes, distance from it.

Similarly to prior work we compile a corpus of DVD subtitles in order to perform in-domain subtitle translations. We train our own SMT model and compare it against other MT approaches and a TM. Our main goal is to demonstrate that, regardless of the MT/TM strategy, PE is faster than HT without a loss in quality. For that, we design a comprehensive evaluation: i) objectively in terms of time (Specia, 2011), ii) subjectively using well specified scoring guidelines (Specia, 2011), and iii) automatically in terms of BLEU using single and multiple references. As a by-product, a comparison between different translation approaches is performed.

3 Corpus

The corpus used in this research was compiled with subtitles from the American TV series "*X Files*" which were downloaded from the free subtitle websites "TVsubtitles.net"[1] , "All-subs.org"[2] and "Opensubtitles.org"[3], where fans of the series volunteer to transcribe and translate subtitles. The corpus presented several types of noise which had to be cleaned such as: i) spelling errors, ii) non-uniform character casing, iii) different encoding, and iv) XML-like tags.

Subsequently, the corpus was automatically aligned at the sentence level using heuristics aimed at maximizing the time overlap between the source and target subtitles. The sentence alignment was revised to guarantee the largest possible set of 1-1 correspondences and also to correct mistakes that resulted from the particularities of aligning subtitles. After the correction of the sentence alignment, four episodes were randomly chosen and kept aside as our *test data*. Statistics about the resulting sentence-aligned parallel corpus are reported in Table 1.

Corpus	Training	Test
`en` tokens	720,845	17,796
`pt` tokens	613,201	14,000
Sentence pairs	76,295	2,379

Table 1: Token and sentence numbers in the parallel corpus

4 Experiments

This section describes how the effort to translate subtitles from scratch was compared to the effort to post-edit translations automatically obtained through different tools.

4.1 Systems

We used three translation tools in this research: two MT systems and a TM system:

RBMT: we used the commercial RBMT system Systran SMTU[4] as a black-box tool.

TM: we used the TM system Trados Studio[5] with a translation memory built using the parallel corpus described in Section 3. To restrict human intervention at the PE stage, we used the *auto-translate* option available in the toolkit. This option ensures that all 100% source matches are automatically translated. As for

[1] http://www.tvsubtitles.net/
[2] http://www.allsubs.org/
[3] http://www.opensubtitles.org/
[4] http://wwwv5.systransoft.com/
[5] http://www.trados.com/en/sdl-trados/default.asp

the remaining segments, the first match retrieved respecting a 70% fuzzy match threshold is accepted without manual correction. When no match is found, the original sentence is copied in the output.

In-domain SMT: we used the parallel corpus (Section 3) to train an `en-pt` phrase-based SMT system using the Moses toolkit (Koehn et al., 2007). The training set was further divided into 74,295 sentence pairs for phrase extraction and the remaining 2,000 sentences pairs for tuning the parameters of the system. For language modeling, we used the Portuguese side of the parallel corpus, along with 262K additional out-of-domain sentences from the Lácio-Ref corpus (Aluisio et al., 2003).

Out-of-domain SMT: we used the SMT system Google Translate as a freely available wide-coverage black-box tool.

4.2 Post-editing Task

Eleven volunteers participated in our experiments: they are native speakers of Brazilian Portuguese and fluent speakers of English and have some experience with translation tasks. They were sent guidelines and asked to post-edit automatic Portuguese translations and to translate English subtitles from scratch.

In order to anticipate any problems the translators could have with both the tool's interface and the task guidelines, and to calculate the translators' agreement regarding the subjective PE assessments (Figure 1), a pilot test was performed. Six translators participated in the pilot test which lasted one week. Each translator post-edited and evaluated the same set of 30 sentences with 10 sentences repeated for intra-agreement computation. Using the Kappa index (Landis and Koch, 1977), an average inter-agreement rate of 0.48 (moderate) and an average intra-agreement rate of 0.69 (substantial) were obtained.

The main experiment was set to last two weeks (W1 and W2) and the translators were divided into two groups (G1 and G2). In W1, 125 English subtitles (sources) were randomly selected from the test set. For every source we produced 4 automatic translations (using Google, Systran, Moses and Trados) which were post-edited by every member of G1. At the same time, members of G2 translated the 125 original source sentences without the aid of any of the translation tools (they could use dictionaries, but no translation tools).

To prevent any bias in the time measurement towards HT or PE, G1 and G2 performed different tasks (translation or post-editing) in the experiment with the same test (source) sentences, and we never asked the same translator to post-edit the output of a source sentence that he/she had previously translated or vice-versa.

Since we were also interested in collecting evidence to compare the effort on post-editing the output of different MT/TM systems, we used the same PE task for pairwise system comparisons. For every source we combined the 4 systems' outputs in pairs, resulting in 6 pairs that were randomly assigned to the members of G1. To avoid assigning more than one comparison pair to a given translator, we had 6 translators performing the PE task. It is worth highlighting that the jobs were distributed during the week, so we could randomly distribute the two automatic translations being compared on different days, reducing the chances that a translator would notice the presence of source duplicates.

In W2 we selected another 125 source sentences and repeated the process swapping the roles of translators in G1 and G2. The purpose of having two weeks and swapping the roles of the groups was to gather effort indicators on HT and PE from the same human translators. Because there were 6 system combinations, the group performing the PE tasks in W2 also had to contain 6 translators. Since we only had 11 translators, one translator did not participate in the HT task and participated twice in the PE task.

We implemented a simple tool to aid the translators performing both tasks. The tool presents the source sentence and its recent context and, in the case of the PE task, the automatic translation. After the translation or post-editing of a sentence, the tool queries the translator for an assessment of the effort put into translating/post-editing the sentence. For the PE task, the translator answers the question '*How much post-editing effort did the translation require?*' and for the HT task, '*How hard was it to translate the source text?*'. The scales for PE and HT assessments are shown in Figures 1 and 2, respectively. Clear guidelines explaining these options were given to the translators.

Score	Description
1	Complete retranslation
2	A lot of post-editing but quicker than translation
3	A little post-editing
4	No modification performed

Figure 1: Scale for PE evaluation

Score	Description
1	Difficult
2	Moderate
3	Easy

Figure 2: Scale for translation evaluation

The PE tool logs the time spent to translate or post-edit individual sentences. Translators can therefore pause between sentences, but they were asked to avoid pausing when possible. Translators were asked to translate/post-edit the sentence literally when it lacked context. Additionally, for post-editing, they were asked to perform the minimum amount of editing necessary to make the translation ready for publishing.

5 Results

To compare different translation tools, we used the human assessments for PE effort collected using the PE tool, as well as BLEU, a standard automatic evaluation metric, computed here for the draft translations before their post-editing. We computed BLEU i) using a single reference translation, that is, the original fan-sub subtitles in Portuguese (ref_0) and ii) using multiple references collected as part of the HT/PE task (i.e. ref_0, five translations made from scratch ref_{1-5} and twelve post-edited translations ref_{6-17}). The aim was to measure how close to any manually obtained translation the MT and TM outputs were and what percentage of the draft translations was reutilized in the PE task. Table 2 compares the performance of the four systems according to BLEU.

References	Google	Moses	Systran	Trados
Single	21.51	**22.28**	13.90	09.22
Multiple	**92.24**	72.04	70.23	28.36

Table 2: BLEU scores using single and multiple (18) references

Overall, both SMT systems outperform the RBMT and TM tools. By comparing the scores one can observe that when BLEU is computed with ref_0 only, Moses has a slightly better performance than Google, even though Google is certainly trained using much larger corpora. This may be due to the fact that Moses was trained using in-domain data, i.e., the corpus with subtitles of the same series. As a consequence, it is more likely that Moses learns specific vocabulary from the series and that translations look more similar to those in the reference set. However, when BLEU is computed with multiple references, even though the translations from all systems may differ from what was originally expected (ref_0), they can still be valid alternative translations that often match the choices made by other translators (ref_{1-17}). This resulted in a different ranking where Google significantly outperforms all other systems. While Moses and Systran have very similar scores, the TM system still performs poorly.

It is worth noticing that TM systems are not meant to be used without human intervention, and therefore our settings tend to penalise Trados, particularly in terms of lexical matching metrics such as BLEU. In fact, unless a full match is possible, all options produced by the TM will contain some noise or words in the source language. Table 3 illustrates the percentage of matches of different types retrieved by Trados. Although BLEU is certainly not a good metric for Trados, it is interesting to compare the TM with Moses, since both are based on the same parallel corpus.

Test set	Full	Fuzzy	Untranslated
Average	1.79%	58.55%	38.66%

Table 3: Different types of matches retrieved by Trados with a 70% threshold for fuzzy matches

In addition to BLEU, the subjective human assessments for PE effort were also compiled. Table 4 shows the percentage of translations assigned different effort scores. More than 92% of the sentences translated by Google were scored as no or little post-editing needed (scores 3 and 4). Over 70% of Moses' and Systran's outputs were also scored 3 or 4. Trados required little or no post-editing for only 36% of its outputs. The MT systems had no more than 8% of the sentences requiring complete retranslation. Trados, however, had more than 47% of its outputs scored as 1. These results are very well aligned to those in Table 2, confirming the BLEU scores using multiple references.

System	1	2	3	4
Google	1.73%	6.00%	28.80%	63.47%
Moses	4.27%	18.40%	36.80%	40.53%
Systran	7.47%	17.73%	40.40%	34.40%
Trados	47.47%	15.87%	19.20%	17.47%

Table 4: How often post-editing a system output was scored 1, 2, 3 or 4

The comparison of translation tools according to the time needed to post-edit their outputs shows that the statistical systems produce translations that require less time to be post-edited. Table 5 illustrates the system comparison in terms of PE time.

System	Google	Moses	Systran	Trados
Google	-	139	161	187
Moses	69	-	122	164
Systran	69	106	-	145
Trados	48	67	89	-

Table 5: How many times the system in the first column produced an output that was more quickly post-edited than each of the other systems (other columns)

According to these time measurements, Google seems to produce the most outputs which can be post-edited in less time as compared to all other systems. Out of 250 cases, Moses was faster to post-edit than Google on 69 translations, while Google was faster than Moses on 139 translations. Although Moses seems to perform slightly better than Systran, both systems are very close: i) both were faster than Google on 69 sentences, ii) Moses was faster than Systran on 122 sentences against 106 for the rule-based, and finally iii) both outperform Trados.

When the systems are compared regarding PE effort assessments, as shown in Table 6, the results are similar to those using PE time, demonstrating a good correlation between objective and subjective effort indicators.

System	Google	Moses	Systran	Trados
Google	-	97	115	186
Moses	22	-	73	162
Systran	30	65	-	159
Trados	8	11	40	-

Table 6: How many times the system in the first column produced an output that was better scored than each of the other systems

To support our main claim in this paper that post-editing draft translations requires less effort than translating text from scratch, we compared the PE effort and HT effort in terms of time. Table 7 shows that post-editing the output of any system is faster than translating subtitles from scratch.

System	Faster than HT
Google	94%
Moses	86.8%
Systran	81.20%
Trados	72.40%

Table 7: How often post-editing a translation tool output is faster than translating the text from scratch

While Table 7 shows how frequently PE is faster than HT, Table 8 shows the actual difference in time. By comparing the average time each translator spent on translating and to post-editing sentences we reach an average ratio (PE/HT) of 0.5952 with a $\pm$0.098 standard deviation, that is, the time to perform PE represents on average about 60% of the time to perform HT. The small standard deviation supports the assumption that PE is 40% faster than HT, regardless of the translator and the source of automatic translations. In other words, translating from scratch consistently takes 70% longer (HT/PE) than post-editing the same sentence.

Annotator	HT (s)	PE (s)	HT/PE	PE/HT
Average	31.89	18.82	1.73	0.59
Deviation	9.99	6.79	0.26	0.09

Table 8: Comparing the time to translate from scratch (HT) with the time to post-edit MT (PE), in seconds

As an additional experiment to study the relation between sentence length and PE effort in terms of time and subjective scores, in Tables 9 and 10 we analyzed the data according to different categories of PE and HT effort scores. Table 9 summarizes the percentage of outputs scored 1-4, the average source length and the average time spent on post-editing, including standard deviation. Table 10 summarizes the same aspects for the sentences translated from scratch.

Score	Samples	Time		Length
		Average (s)	Deviation	
1	15.2%	32.19	29.95	8.503
2	14.0%	40.87	50.98	9.343
3	31.0%	18.92	20.63	7.924
4	38.9%	5.02	8.15	6.122

Table 9: Correlation between PE effort score and average input sentence length

Score	Samples	Time		Length
		Average (s)	Deviation	
1	7.04%	111.19	82.875	10
2	18.96%	53.21	38.875	9
3	74.0%	20.29	19.342	6.89

Table 10: Correlation between translation effort score and average input sentence length

In Table 9 we can see that sentences scored 4 took on average 5 seconds to be post-edited. This may be because the tool did not permit the translators to read a sentence before they started post-editing it, thus 5 seconds would be the average time the translators spent reading the source sentence and its suggested translation, to then decide that it did not needed any post-editing.

More than 38% of the sentences were scored 4 (no modification performed) and more than 69% were designated as little or no post-editing performed. Although Tables 9 and 10 provides a certain pattern regarding the length of sentences and the scores (shorter sentences seem to have higher scores); it is interesting to note that sentences scored 1 are surprisingly shorter than sentences scored 2. Our hypothesis is that sentences that are shorter and contain several errors are more likely to be deleted whereas longer sentences tend to be fixed because it saves time on typing. It seems to take less effort to erase and rewrite short sentences than to reorder them.

It is worth noticing that post-edited translations scored 1-2 in Table 9 and sentences translated from scratch scored 2 in Table 10 have a similar length, which allows us to compare them in terms of time. Table 9 shows that post-editing sentences scored 2 is a bit slower than sentences scored 1 (requires complete retranslation). Nevertheless it does not mean that post-editing those sentences is slower than translating their original sources from scratch. We can see in Table 9 that post-editing a sentence that requires complete retranslation (scores 1-2) is less time-consuming than translating the same sentence from scratch (score 2 in Table 10). This may be so because even when the sentence requires complete retranslation the translator may benefit from the translation of some terms even if he or she considers the translation inappropriate for the sentence. The output sentence may provide the translator with a gist of the translation whereas translating from scratch also involves the effort of considering several possibilities for translating the source.

Finally, we were concerned with the quality of the post-edited translations. Although the translators were asked to perform the minimum necessary operations while post-editing, they were instructed to produce translations that were "ready for publishing". We conducted an automatic evaluation comparing each of the 12 sets of post-edited translations to the 5 sets of translations made from scratch and the corpus-based reference (ref_{0-5}). We observed a high average BLEU score of 69.92 $\pm$ 4.86 (less than 7% standard deviation), which suggests that PE does not imply any loss in translation quality, as compared to standard translations. It is always important to highlight that post-edited translations that do not match a reference are not necessarily bad as they could still be valid paraphrases. A human evaluation of these aspects is yet to be performed.

6 Conclusions and Future Work

We presented experiments showing that automatic and semi-automatic translation of DVD subtitles may be of great help to subtitlers, since the pre-translated subtitles are proven to be less time-consuming to post-edit than translating from scratch. As expected, we found a high correlation between a subjective scoring of the post-editing effort and the actual time necessary to post-edit translations. In addition, we found a strong correlation between this scoring and sentence length: high scoring translations are usually those with short length. Nevertheless, Table 10 gives us an insight that short sentences that contain several errors are more likely to be completely discarded and translated from scratch.

Regarding the performance of the translators, Table 8 confirms that the average time spent to translate from scratch is more than 70% higher than the time to post-edit the same sentence. Further analysis has shown that all the translators had a better time performance when post-editing a pre-translated sentence. The number of times that PE was faster than HT (Table 7) is substantial proof

of our hypothesis. Even the TM system, which often did not perform as well as the other MT systems, achieved a high performance when compared against translating from scratch. Translating with TM systems may be a way of ensuring consistency in the translation, that is, the TM system may help the translator to be consistent when translating the same sentence more than once.

We believe that by treating punctuation and character case and by having a larger corpus, the TM system would retrieve a greater number of high percentage matches. A larger corpus would obviously contribute to a better performance of the SMT Moses as well. The rule-based system could also have an improved performance if its linguistic resources were specific to the subtitle domain, maybe by extracting in-domain bilingual dictionaries from parallel corpora.

Despite the small size of the corpus, it became evident that automatic and semi-automatic translation of subtitles can be a real help for subtitlers by speeding up the translation process by 40% for most of the subtitles (from 72 to 94% depending on the translation engine). This can also mean in practical terms a cost reduction for subtitling companies.

In future work, to clarify some choices regarding scores the translators have made, a questionnaire will be developed in order to have a more detailed analysis of the output of the systems. We also want to evaluate the subtitles including the process of fitting the translation according to specific restrictions in the field: time and length. In addition, the post-edited subtitles could be evaluated by native speakers of the target language (regarding quality) in the role of real end-users watching the videos with subtitles.

Acknowledgments

This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme. We would like to thank Professor Dr. Jorge Baptista for his support and insights and all translators who participated in this experiment.

References

Sandra M. Aluisio, Gisele Pinheiro, Marcelo Finger, Maria G. V. Nunes, and Stella E. Tagnin. 2003. The lacio-web project: overview and issues in brazilian portuguese corpora creation. In *Corpus Linguistics*, pages 14–21, Lancaster, UK.

Stephen Armstrong, Colm Caffrey, and Marian Flanagan. 2006. Translating dvd subtitle from english-german and enlish-japanese using example-based machine translation. In *Audiovisual Translation Scenarios*, MuTra '06, pages 1–12, Copenhagen, Denmark.

Marian Flanagan. 2009. Using example-based machine translation to translate dvd subtitles. In *Proceedings of the 3rd Workshop on Example-Based Machine Translation*, pages 85–92, Dublin, Ireland.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL: Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Maite Melero, Antoni Oliver, and Toni Badia. 2006. Automatic multilingual subtitling in the etitle project. In *Proceedings of the Twenty-eighth International Conference on Translating and the Computer*, Aslib '06, London, UK.

Minako O'Hagan. 2003. Can languge technology respond to the subtitler's dilemma? - a preliminary study. In *Proceeding of the 25th International Conference on Translation and the Computer*, London, UK.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Fred Popowich, Paul Mcfetridge, Davide Turcato, and Janine Toole. 2000. Machine translation of closed captions. *Machine Translation*, 15:311–341.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Annual Conference of the European Association for Machine Translation*, EAMT '11, Leuven, Belgium.

Martin Volk. 2008. The automatic translation of film subtitles. a machine translation success story? In *Resourceful Language Technology: Festschrift in Honor of Anna*, volume 7, Uppsala, Sweden.