An Analysis (and an Annotated Corpus) of User Responses to Machine Translation Output

Daniele Pighin[†], **Lluís Màrquez**[†], **Jonathan May**[‡]

[†]Universitat Politècnica de Catalunya, Barcelona
[‡]SDL Language Weaver
{pighin,marquez}@lsi.upc.edu, jmay@sdl.com

Abstract

We present an annotated resource consisting of open-domain translation requests, automatic translations and user-provided corrections collected from casual users of the translation portal http://reverso.net. The layers of annotation provide: 1) quality assessments for 830 correction suggestions for translations into English, at the segment level, and 2) 814 usefulness assessments for English-Spanish and English-French translation suggestions, a suggestion being useful if it contains at least local clues that can be used to improve translation quality. We also discuss the results of our preliminary experiments concerning 1) the development of an automatic filter to separate useful from non-useful feedback, and 2) the incorporation in the machine translation pipeline of bilingual phrases extracted from the suggestions. The annotated data, available for download from ftp://mi.eng.cam.ac.uk/data/faust/LW-UPC-Oct11-FAUST-feedback-annotation.tgz, is released under a Creative Commons license. To our best knowledge, this is the first resource of this kind that has ever been made publicly available.

Keywords: Machine Translation; Feedback Filtering; Annotated Corpus

1. Introduction

The *Feedback Analysis for User adaptive Statistical Translation* (FAUST) EU project[1] focuses on the development of machine translation systems that can respond rapidly and intelligently to user feedback. As such, it is centered around user provided translation requests and the responses of users to machine translation output. Within this project, we carried out an analysis and annotation of a corpus of open-domain, real-world automatic translations together with the correction suggestions left by the users of an online translation service. The web service is run by *Softissimo*[2], which relies on *Language Weaver*[3] technology to actually satisfy the translation requests[4]. The feedback data consists of quality ratings, suggested translations and comments about the service, on which we carried out two annotation activities with different methodologies. The annotations, available for download from ftp://mi.eng.cam.ac.uk/data/faust/LW-UPC-Oct11-FAUST-feedback-annotation.tgz, focus on the comparison of automatic vs. user provided translations, and are aimed at understanding if and how it is possible to characterize suggestions that can be leveraged to improve the output of machine translation. To our best knowledge, this is the first resource of this kind that has ever been made publicly available.

The annotated resource can be used to improve machine translation (MT) systems by discovering common pitfalls and usual corrections to typical translation errors. But more then that, the data tells a story about the difference between real-world usage scenarios and user expectations, which appear to be remarkably far from the controlled environments (Callison-Burch et al., 2009; Callison-Burch et al., 2010) in which state-of-the-art MT models (Koehn et al., 2007; Chiang et al., 2005) are generally conceived, developed and carefully tuned. In this respect, the annotated resource constitutes a valuable asset towards the development of user-centered, adaptive machine translation technology.

The rest of this paper is organized as follows: in Section 2 we will present an overview of the data used for our analysis, and a first annotation of 830 pairs consisting of automatic and user-edited English translations; in Section 3 we discuss the annotation of user-suggestions as useful or not-useful, depending on the presence of sub-sentential clues that may be used to improve translation quality, and we document a first attempt to implement an automatic user-suggestion filter based on these annotations; in Section 4 we show how the useful feedback can be exploited to improve the accuracy of MT output; finally, in Section 5 we will draw our conclusions.

2. Feedback Analysis

The FAUST project is collecting user feedback through *Softissimo*'s web portal, which serves approximately 30 million bi-directional translations per month between English and nine other languages. Users of the service can input the text they want to translate in a text box, select the direction of the translation and optionally ask for the text to be spell-checked prior to translation. After reading the translation, users can assign it a grade between 1 (barely comprehensible) and 4 (fully comprehensible). If they do so, a window pops up asking for the user to add a comment about the translation and to provide a better alternative. This step is optional, and either or both text boxes can

[1] http://www.faust-fp7.eu/faust

[2] http://www.reverso.net

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

[4] Both *Softissimo* and *Language Weaver* are partners of the FAUST consortium.

in fact be left empty. Hence, each feedback item is a tuple consisting of:

- the language pair selected in the translation interface;
- the source sentence;
- the automatic translation;
- the rating given by the user;
- the (possibly empty) commentary by the user; and
- the (possibly empty) correction by the user.

A first datum to point out is that fewer than 0.01% of translation requests receive any kind of feedback. Furthermore, most users limit their interaction to selecting a rating, and hardly leave any comment or suggest an alternative translation.

As a basis for our analysis, we selected a body of 50,000 user feedback forms. After filtering segments longer than 30 words and non-empty feedback forms, we were left with only 11,779 items, most of which are relative to translations from English to French (2,904), English to Spanish (2,509), French to English (1,652) and Spanish to English (1,389).

A preliminary analysis of the data, carried out by sampling 100 feedback forms, shows that feedback provided by users generally falls under the following categories:

Appreciation: 32% of the users express their gratitude for having access to the translation service;

Criticism: 15% of the users are disappointed by the translation engine;

Non-Useful feedback: 32% of the feedback is either non-informative, completely unrelated, a worse translation than the automatic one, or any other text that is difficult to see how to exploit computationally;

Useful/Good feedback: 21% of the feedback contains useful hints to improve MT output, either in the form of a better translation or as a remark about translation errors, e.g., "The conjugation of the verb is wrong.".

While all types of feedback may prove useful, for the annotation we focus on translation corrections as this feedback is more regular and reliable than open-ended commentary, it has the potential to be more informative than quality ratings, and it can more easily be exploited computationally, e.g., by extracting correct phrases to be added to existing translation tables.

Feedback validation

To further investigate the quality of feedback, we set up an annotation activity with the intention to compare automatic translations against user-provided suggestions. The activity involved 19 annotators who annotated 830 automatic/user-provided English translations, regardless of the source language. For each pair, the annotators answer from 1 to 4 yes/no questions, based on the traversal of a decision tree. At the end of this series of questions, each automatic/suggested translation pair is classified into one of five categories:

Decision	% Cases
% Semantically opaque	28.64
% Poor suggestion	43.32
% Automatic is good	0.48
% Both are good	4.09
% Suggestion is good	13.36
% No decision taken	9.99
% Strictly good	17.45

Table 1: Results of the annotation of well-formed suggestions.

- The suggestion provided by the user is unreliable, i.e., disfluent, ungrammatical or non interpretable (*Poor suggestion*)
- It is not possible to assess the semantic equivalence between the automatic and the user provided translation (*Semantically opaque*);
- The automatic translation is fluent and grammatical, and better than the suggestion (*Automatic is good*);
- The user-provided translation is fluent and grammatical, and better than the automatic one (*Suggestion is good*);
- Both the automatic and the user provided translations are grammatical, and convey the same amount of information (*Both good*).

Of these five classes, the last two correspond to *Strictly good* feedback. It should be noted that disregarding the source text during the annotation may lead to wrongly discard good suggestions, e.g., a good suggestion may be discarded as *Semantically Opaque* due to an especially poor automatic translation. On the other hand, this annotation procedure requires linguistic competence in only one language, and it focuses on the quality of the output rather than on the comparison between input and output. The simplified setting seems to enforce annotation consistency, as in 90.12% of the cases the annotators take the same decision when presented with the same pair. Concerning inter-annotator agreement, we measured Cohen's κ Cohen (1960) between the two most productive annotators. On a set of 101 shared annotations, the probability of agreement between the two annotators is 0.75, and The value of the coefficient is $\kappa = 0.61$.

The idea here is that a suggested translation is especially useful for improving MT output if it is as close as possible to the automatic translation and if it does not contain errors, which could be difficult to identify and isolate. While there is no direct evidence of the adequacy of the translation with respect to the source sentence, the semantic convergence of the automatic and the user provided translations are a strong indicator of the fact that the automatic translation has been perceived as adequate by the user who suggested the correction.

A summary of the resulting annotations are shown in Table 1. The top block of rows shows the percentage of items

Language Pair	Annotations done	% Useful
English→Spanish	245	60.0
English→French	569	63.4

Table 3: Results of the annotation of useful feedback.

classified into each class, according to the class selected by the relative majority of annotators. The row labeled *No decision taken* accounts for all the cases in which no class can be selected. The row labeled *Strictly good* aggregates the results of *Both are good* and *Suggestion is good*. These figures suggest that exploiting user-provided translations at the passage level may be very difficult, as only in 17.45% of the cases the suggestion as a whole is equivalent to or better than the automatic translation, while in more than 43% of the cases the user-provided translation is actually worse than MT output. This effect is somehow expected, as it is reasonable to assume that many users of an online translation portal are not fluent speakers of either the source or the target language. In the light of these results, we carried out another annotation activity with the aim of isolating user suggestions that may contain useful sub-sentential clues to improve translation quality, as described in the next section.

3. Recognizing Useful Suggestions

One professional translator labeled 569 French→English and 245 Spanish→English correction entries as *Useful* or *Not Useful*, based on clues in the correction that may be employed to improve translation quality.

The annotator was shown triplets of passages in which the automatic and the user-provided translation are explicitly marked and shown next to the source passage. A triplet was classified as useful if the correction is a better translation of the source than the machine translation. The suggestion needs not be perfect in order for the triplet to be marked as useful, it just needs to be an improvement over the automatic translation. Anything else, such as a correction that makes the translation worse, commentary mistakenly added to the correction box, junk, or badly formed data, was judged not useful. Those cases in which there is no difference between the suggestion and the automatic translation were also marked as not useful. The annotator was not required to adopt a specific quality criterion. Table 2 shows two examples of useful and not useful entries from the English→Spanish dataset.

The results of this annotation are shown in Table 3. The table shows that approximately sixty percent of the feedback items contain at least a partial improvement over the automatic translation.

Automatic feedback filtering

We used the annotations to learn a feedback classifier to determine whether a (Source, MT, Correction) feedback triplet is useful or not. To build our classifier we identified a set of likely helpful feature classes, extracted features for the entries, then divided the annotated data into training and test sets. As a learning framework, we used the Maximum-Entropy model optimizer MegaM [5]. We experimented with the following feature classes:

Surface features: calculated from the source sentence s, the automatic translation t and the post-edited translation (correction) p:

- word/character Levenshtein distance between p and t, divided by the length of t;
- *p*/*t* word overlap, divided by translation length;
- translation words not in correction, divided by translation length;
- correction words not in translation, divided by correction length;
- *p*/*t* word overlap with the source, divided by the length of s;
- length of *s*/*t*/*p*;
- length of *t*/*p* divided by length of s;
- average/maximum length of words in *s*/*t*/*p*;
- average/maximum length of words in the *t*/*p*, divided by average/maximum word length in s;
- average/maximum length of words in t, divided by average/maximum word length in p;
- a binary feature indicating if p and s are the same string.

Back-translation features: calculated by first generating back-translations t' and p' for t and p, respectively, into the source language:

- word/character Levenshtein distance between t'/p' and s;
- word/character Levenshtein distance between t' and p';
- t'/p' word overlap with s, divided by source length;
- number of words in t'/p' which are not in s, divided by the length of t'.

We compare this approach to the baseline approach of simply selecting all correction entries as useful. Since the extraction of BT features relies on the creation of back-translations of translated and corrected sentences, and is thus more costly than simply using surface features, we used scenarios both with and without these extra features, to determine if their inclusion justifies the extra work. As the training sets for each language pair are fairly small, as well as to determine if the filtering technique is language-independent, we learned different models using only English-Spanish data, only English-French data, or by combining the two training sets. We tested each of these systems on held-out English-Spanish and English-French test sets.

[5] http://www.cs.utah.edu/~hal/megam/

Source	Auto	Suggestion	Comment
it's gonna be alright	esto va a ser bien	esto va a estar bien	Fixes verb
I'm writing to tell you about the party last night	Escribo para decirle sobre el partido anoche	Escribo para decirle sobre la fiesta de anoche	Fixes bad translation of "party"
goat's milk	la leche de la cabra	leche de la paja	Auto was better
bobby always does his homework	Bobby siempre hace su tarea	hacerme caca	Non-responsive

Table 2: Examples of useful (top) vs. non-useful (bottom) suggestions.

		Test								
		EN-ES (61)			EN-FR (142)			Macro-AVG (203)		
Setup	Training	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc
Baseline	n/a	55.7	100	55.7	65.5	100	65.5	62.6	100	62.6
Surface features	EN-ES (184)	56.8	85.3	55.7	69.0	93.5	68.3	65.3	91.0	64.5
	EN-FR (427)	63.4	76.5	62.3	75.5	86.0	72.5	71.9	83.1	69.4
	both (611)	60.0	79.4	59.0	75.5	89.2	73.9	70.8	86.3	69.4
Surface + BT features	EN-ES (184)	63.0	85.3	63.9	71.8	95.7	72.5	69.2	92.6	69.9
	EN-FR (427)	64.1	73.5	62.3	74.5	78.5	68.3	71.4	77.0	66.5
	both (611)	64.3	79.4	63.9	73.9	88.2	71.8	71.0	85.6	69.4

Table 4: Feedback filtering accuracy.

The results of these experiments are shown in Table 4, where we report classification accuracy (i.e., the percent of samples classified correctly), precision (i.e., the percent of sentences classified useful that actually are) and recall (i.e., the percent of useful sentences that are so classified) of the different classifiers. In order to identify relevant clues to improve MT output (see next section) our focus here is especially on precision, as the availability of large quantities of feedback makes it possible to compensate for a lack in recall. Precision and accuracy-wise, the feedback classifier performs consistently better than the baseline. While the best results can be achieved using French training data to classify French data, using a mixed training set results yields a more accurate classifier, suggesting that the task is fairly language-independent and that it could benefit from the availability of more training data. This is especially evident when also back-translation (BT) features are included. In this case, mixing training data provides a boost in recall while leaving precision almost unaffected. That said, more data and more sophisticated features may be needed in order to implement a reliable classifier, as even the best configuration that we experimented yields a maximum precision of 75.5%.

4. Exploiting Suggestions to Improve SMT

After selecting good feedback entries, we are faced with the challenge of determining sub-sentential dictionary entries motivated by the feedback that we may employ to improve translation accuracy. We specifically seek phrase pairs that are implied by the feedback entries but are missing (or insufficiently weighted) in an existant MT system's phrase table. Our methodology for extracting dictionary entries from a (Source, MT, Correction) feedback entry corpus, which relies on extant technology for building phrase-based SMT systems, follows these two steps:

Step 1: Construct one phrase table from a (source, translation) bitext and a second phrase table from a (source, correction) bitext:

- Segment words, tokenize, and decapitalize both sides of the bitext;
- Align the words in the bitext using, e.g., an instantiation of IBM models 1 to 4 (such as is available in GIZA++);
- Extract phrase pairs from the bitext that are consistent with the alignments and subject to typical restrictions (e.g. phrase length, unaligned word restrictions).

Step 2: Identify phrase pairs from the two tables that are likely good dictionary corrections, subject to the following restrictions:

- Only phrase pairs with 3 or more words in either phrase are considered;
- Only phrase pairs with terminal words aligned are considered;
- For the considered phrase pairs, for a given source side, if the (source, translation) and (source, correction) phrase tables do not share a target side, and the

Setup	Feedback entries	Dict Entries	%Good Entries	% Phrases from Dict	BLEU (2-ref)
Baseline (no dict)	N/A	N/A	N/A	N/A	23.74
Filtered Feedback	1,749	10,286	56	0.4	23.82
Unfiltered Feedback	2,382	10,677	46	0.3	23.84

Table 5: Effect of user feedback on translation quality.

Setup	Feedback entries	Dict Entries	%Good Entries	% Phrases from Dict	BLEU (1-ref)
Baseline (no dict)	N/A	N/A	N/A	N/A	43.31
Clean Feedback	3,000	43,381	76	6.0	43.91

Table 6: Effect of high-quality, in-domain feedback on translation quality.

(source, correction) table has exactly one target side for that source, the phrase from the (source, correction) table is taken as an entry.

We used this method to extract dictionary entries from the filtered noisy feedback acquired from Softissimo logs. Following the methodology described in Section 3, we trained a feedback classifier on all 814 English-Spanish and English-French annotated feedback entries, using baseline and back- translation features. We then ran this classifier on 2,382 English-Spanish entries containing correction feedback, which we extracted from http://reverso.net feedback logs. The classifier filtered 1,749 of the entries as useful. From those entries, we built a dictionary, as described above, containing 10,286 entries. By way of contrast we also built a dictionary of 10,677 entries from the entire unfiltered 2,382-entry set, to determine the effects of filtering on dictionary quality.

Table 5 shows the evaluation of two MT systems making use of the dictionaries, along with a baseline which does not use this information. For tuning and testing, we use a corpus of 2,000 English-Spanish automatic web-log translations for which two professional translators provided reference translations. We report the average mixed-case BLEU on the two references. We also note the percent of phrases used in the dictionary-enhanced translations that come from the dictionary, and the percent of the dictionary entries that are good-quality, which we judged by annotating randomly sampled entries. The effect of dictionaries here seems to be marginal, as well as the effect of filtering. This is likely due to the relatively low precision of the useful feedback classifier, the small amount of feedback available for dictionary extraction, and the consequent low quality of the entries that were extracted. Still, the very low percentage of phrases coming from the dictionaries (∼.3%) seem to have a positive effect on translation quality.

Hypothesizing that good-quality feedback leads to good-quality dictionary entries and translations, we repeated the dictionary extraction experiment, but this time extracted entries from a corpus of 3,000 professionally post-edited feedback entries in the English-Spanish technical manual domain. Using the above methodology we extracted 43,481 entries. We translated a held-out test corpus of 2,839 sentences using both our baseline and MT enhanced with the dictionary. In Table 6 we show the effects of using these entries on single-reference, mixed-case BLEU. These results indicate that a modest gain can be obtained using clean feedback and a heuristic method for dictionary extraction. Clean, in-domain feedback leads to a dictionary set with more good-quality entries that is used more often during decoding.

5. Conclusions

In this paper we have described our analysis and the annotation activities around a corpus of open-domain, user-provided translation requests and the corresponding translation corrections gathered through the web portal http://reverso.net. We have discussed two annotation activities: the first, aimed at characterizing the usefulness of user-suggestions at the segment level, showed that it is very difficult for casual users to produce adequate and fluent translations; the second, aimed at identifying useful clues to improve translation accuracy in the feedback, suggests that most feedback items contain at least some form of local improvement that we should try to exploit.

Only a very small fraction of translation requests receive any kind of feedback from users (∼0.01%), and after filtering out noisy feedback items, approximately 60% of the suggestions contain at least some clue on how to improve translations, while only ∼17% are completely adequate and correct. These findings clearly show that collecting useful feedback from users is a challenging task and baseline methods are not well-suited to adequate data collection.

We have shown that the annotated data can be leveraged to learn a classifier to automatically separate useful and non-useful feedback with a precision of 75%. While this figure can certainly be improved, we have demonstrated that the filter can be used to extract phrase-pairs to be used to improve, at least marginally, the quality of MT output. As shown in another experiment, larger amounts of cleaner, possibly in-domain data are needed in order to obtain more noticeable improvements of translation quality.

Despite the difficulties associated with collecting user feedback and the profound noisiness of the data, we are excited to be the first, to our knowledge, to both provide the community with a significant corpus of real-world feedback and to provide annotation schemes for a portion of that corpus. The data, available for download from ftp://mi.eng.cam.ac.uk/data/faust/ /LW-UPC-Oct11-FAUST-feedback-annotation.tgz under a Creative Commons license, provides useful insights about what casual users want to translate, and about their honest reactions to the systems providing the translations.

Acknowledgements

This research has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009-4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4-247914) and by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03).

6. References

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder, editors. 2009. *Proceedings of the Fourth Workshop on Statistical Machine Translation*. ACL, Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. ACL, Uppsala, Sweden.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. 2005. The hiero machine translation system: extensions, evaluation, and analysis. In *Proceedings of HLT'05*, pages 779–786, Vancouver, British Columbia, Canada. ACL.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

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 ACL'07*, pages 177–180, Prague, Czech Republic. ACL.