Using Automatic Machine Translation Metrics to Analyze the Impact of Source Reformulations

Johann Roturier
Symantec Research Labs
Ballycoolin Business Park
Blanchardstown, Dublin 15, Ireland
johann_roturier@symantec.com

Linda Mitchell
SALIS
Dublin City University
Ballymun, Dublin 9, Ireland
linda.mitchell7@mail.dcu.ie

Robert Grabowski, Melanie Siegel
Acrolinx GmbH
Friedrichstr. 100, 10117 Berlin, Germany
{robert.grabowski,melanie.siegel}@acrolinx.com

Abstract

This paper investigates the usefulness of automatic machine translation metrics when analyzing the impact of source reformulations on the quality of machine-translated user generated content. We propose a novel framework to quickly identify rewriting rules which improve or degrade the quality of MT output, by trying to rely on automatic metrics rather than human judgments. We find that this approach allows us to quickly identify overlapping rules between two language pairs (English-French and English-German) and specific cases where the rules' precision could be improved.

1 Introduction

Software publishers rely on manuals and online support (knowledge base) articles to assist their users with the installation, maintenance or troubleshooting of products. With the emergence of Web 2.0 communication channels, however, these documentation sets have been supplemented with User Generated Content (UGC). Users are now extremely active in the generation of content related to software products, especially on online forums, where questions are being asked and links to solutions exchanged among savvy users. While specific language versions of such forums sometimes exist, most content is very often written in English and may require translation to be of any use to specific users. While such content is sometimes machine-translated (e.g. hotel reviews[1]), some comprehension problems may persist. These comprehension problems on the target side may be caused by the following characteristics of UGC on the source side:

- Source content may be written by non-professionals or non-native speakers (so its linguistic and technical accuracy may not be optimal).
- Although written, this content is closer to oral content, with informal syntax and creative lexicon.
- Some of the content is authored by power users who "exhibit communicative techniques that are guided by attitudes of technological elitism (Leblanc, 2005)." These include alternative spellings, acronyms, case

[1] http://blogs.forrester.com/tim_walters/10-07-15-sdl_casts_vote_machine_translation_language_weaver_acquisition

change, techie terms, emoticons, or representation of non-lexical speech sounds.

In this paper, we propose a novel framework to rapidly evaluate the impact of specific reformulations on machine-translation quality. The reformulations provided by this framework are related to the work presented in Section 2. This work is conducted within the ACCEPT project[2], which aims at enabling machine translation for the emerging community content paradigm, allowing citizens across the EU better access to communities in both commercial and non-profit environments.

2 Related Work

Rewriting or reformulating source content to make it more machine-translatable is an active area of research. Several approaches have been used to-date: source normalization, source re-ordering and source control. Our framework provides a way to evaluate the impact of some of these approaches in a rapid manner.

2.1 Source Normalization

Source normalization can be achieved using a number of techniques, including those described in Banerjee et al. (2012): masks using regular expressions, spell-checking and fused word splitting. While these techniques can be effective in reducing OOV words, their impact is somewhat limited in terms of BLEU score improvements when error densities are low.

Another area of sentence normalization involves replacing sentences with similar sentences. Given a large amount of text data in the domain, sentence clustering can find similar sentences and help standardize them. If the variant selection is trained on Machine Translation training data, it can be made sure that in case of variants in the source language text these are changed to 100% matches in the Machine Translation training data. The problem is that User Generated Content is not homogenous so using this approach effectively may prove difficult.

2.2 Source Re-ordering

Another approach is to re-order source text to make it closer to the target text before it gets machine-translated. Such an approach was suggested by Collins et al. (2005) and supplemented by other works, including Genzel (2010). While this approach can produce translation quality improvements (especially in terms of BLEU scores), it is not appropriate when the transformed source text must be published (which might be the case in the context of user-generated content).

2.3 Source Control

Source control (or controlled language), which places restrictions or constraints on lexicon, grammar and style, has been used for a long time in the domain of technical authoring in order to improve the machine-translatability of source text (Bernth and Gdaniec, 2002). Various studies, including O'Brien and Roturier (2007) and Aikawa et al (2007) have since shown that this approach could indeed lead to machine translation quality improvements (either in terms of comprehensibility or post-editing efficiencies).

Since some of the rules are system-, domain- or language-specific, they must be re-evaluated before being used for a new scenario. However, such evaluation can be extremely time-consuming and expensive, especially if two sets of reference translations are required (Doherty, 2012). In this study, we are therefore interested in finding out whether it is possible to quickly identify effective rules, by relying on automatic metrics rather than human judgments.

3 Description of Systems and Data

3.1 Data

The test set used in this paper contains 2031 sentences corresponding to 250 posts that were randomly selected from the English Norton Forum (as described in Banerjee et al. 2012). These sets were then translated by professional translators in order to obtain both French and German reference translations. Professional reviewers were then asked to perform a linguistic and technical review of these translations, with a view to identifying and correcting potential translation errors. The result of this review was used as a second reference translation set.

3.2 MT System

The MT systems used in these experiments are phrase-based Moses systems, trained with the

[2] http://accept-project.eu/

standard Moses pipeline. The translation and reordering models were trained with a concatenation of all the available parallel data, while for the language model a separate model was trained on each corpus, and all models were interpolated together minimizing perplexity on the tuning set. The Moses tokenization and casing tools were used. The parallel data consisted of Symantec's translation memory data (containing product manuals, marketing content, knowledge base content and website content), supplemented with the WMT12[3] releases of Europarl and news-commentary. For the language models, the target sides of all the parallel data were used, together with monolingual data from the Symantec forums. The monolingual data was not included in the English-German system as it was found not to improve the Bleu score. The tuning and test data for the Symantec systems (500 parallel sentences each) consisted of forum data which had been machine-translated with an online MT system and post-edited using the CNLG/TAUS guidelines[4].

3.3 Source Reformulation System

Built on a linguistic analytics engine that provides rules and resources concerning monolingual texts (as described in Bredenkamp et al., 2000), the Acrolinx software provides spelling, grammar, style and terminology checking. These methods of pre-editing can on the one hand be applied by authors, as usually done in the technical documentation authoring process. The author gets error markings and improvement suggestions, and decides about reformulations. This process ensures that text conversions are always correct. Further, a learning process for the author starts. He or she gets a better understanding of the abilities and limits of Machine Translation as such.

On the other hand, it is possible to automatically apply the provided suggestions as reformulations. Different to authoring support of technical documents, the focus here is on better Machine Translation results. Automatic application of rules is much faster. This process only influences the translation, so the precision of the application is not as crucial.

For the purpose of this paper, the Acrolinx software and lingware were adapted to handle User Generated Content, specifically to handle non-native language errors, language close to oral content and language used by "techies". We used the standalone tool called "autoApplyClient" to send documents to an Acrolinx server, retrieve the result, and automatically apply all suggested reformulations by replacing the marked part of the document by its suggestion.

The client has two distinct output modes. In the first mode, it applies all suggestions onto the same document and writes the result into a new (single) document. In the second mode, it applies the suggestions individually: for each possible reformulation, the client outputs the original sentence, the reformulated sentence, the type of error (spelling, grammar, style, or terminology), as well as the name of the applied grammar or style rule, or the preferred variant of the term that has been used. For the evaluation, we used both the global and the sentence-based reformulation mode.

3.4 Scoring using Automatic Metrics

In order to quickly identify pre-editing reformulation types that appear to have an improving or degrading effect on the quality of MT translations, we used automatic metrics to score the translation of the original and of reformulated texts with respect to the reference translations. For the first evaluation of the effects on entire documents, we calculated the following scores: Smoothed BLEU (Lin & Och, 2004) averaged across all sentences, Translation Error Rate (TER as described in Snover et al. (2006) and General Text Matcher (GTM, as described in Melamed et al. 2003), including the precision, recall, and f-measure score[5].

For the second evaluation of the effects on a sentence-level basis, we calculated the following scores per reformulated and original sentence: smoothed BLEU, Translation Error Rate (TER), and General Text Matcher (GTM) f-measure. The "smoothed" modification of BLEU avoids the score becoming zero in case an n-gram does not exist, a common situation in the sentence-based evaluation due to the small reference size. Additionally, we calculated the language model (LM) scores for language models created from the training set used for the translation system. The LM score measures how "similar" a given segment is

[3] http://www.statmt.org/wmt12/
[4] http://www.cngl.ie/node/2542

[5] The GTM scores presented here are based with the exponent set at 1.2, which puts a mild preference towards items with words in the correct order.

to the training set. The LM score is one factor for Moses to determine the most likely translation. Here, we calculated the LM score for both the source and the target language training set, following two hypotheses:

1. a better source language model score makes a text "easier" to translate, because it is more similar to the training corpus,
2. a better target language model score means the text is more similar to a "known-good" corpus of target language sentences and thus has a higher translation quality.

To easily obtain the scoring results, we created a framework that combines the autoApplyClient, the Moses SMT software, and the different automatic scoring metrics as shown in Figure 1:

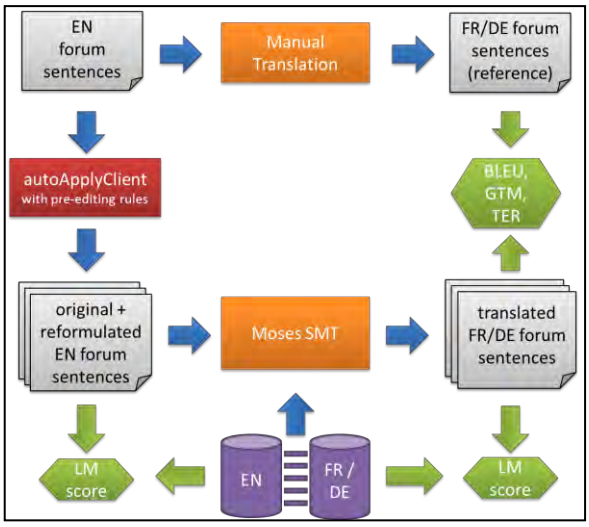

Figure 1: Rule Evaluation Framework

The second experiment focused on scores for each individual reformulated sentence. Since the absolute values are neither comparable among metrics nor among different sentences, we simplified the data set by transforming it to relative scores: for each reformulated sentence, we noted whether a score has improved, degraded, or stayed the same with respect to the score of the corresponding original sentence without reformulations. Note that the "amount" of impact on the scores was not considered.

3.5 Scoring using Human Evaluation

In the second experiment, two factors arguably harm how representative the results are. First, computing scores on the sentence level means that small "misjudgments" by an automatic metric are not leveled out by the size of the data. Second, the switch to a relative better/equal/worse metric means that information about the quantity of the impact of a reformulation is lost.

To counter these effects and get a better understanding of the significance of automatic metrics, we first filtered out all rules that caused no more than 12 reformulations on the given input data. For the reformulations of the remaining rules, we then conducted a human evaluation. The evaluator was given the translation of the original and the reformulated sentences, with the task to judge which translation was closer to the reference translation, or whether there was a change at all. The procedure thus gave another set of scores on better/equal/worse scale. With this "human metric", we were able to examine whether the automatic metrics are consistent with human judgments, before we looked at particular rules that had the most impact based on the automatically computed scores only.

4 Automatic Reformulation Experiments and Results

We conducted two sets of experiments: the first one consisted of automatic reformulations on the whole test set. The second round of experiments examined the scores on the level of individual suggestions.

4.1 Overall Results

Table 1 indicates the number of sentences (out of 2031) changed by the autoApplyClient as well as the amount of difference between the source texts (in terms of TER and GTM F-Measure):

	Sents	TER	GTM F-Measure
Original vs. Spelling	150	0.0101	0.9804
Original vs. Grammar	67	0.0054	0.9901
Original vs. Style	328	0.0334	0.9529
Original vs. Spelling +Grammar	197	0.0157	0.9708
Original vs. Spelling + Grammar + Style	403	0.0483	0.9279

Table 1: Amount of Source Changes

Table 1 shows that grammar reformulations (67) are far less frequent than spelling (150) or style

(328) reformulations. Their impact, however, is more positive, as shown in Table 2.

	Original	Spelling	Grammar	Style	Spell+ Gram	All
FR						
SBLEU*	0.3962	0.3929	0.3974	0.3950	0.3941	0.3931
TER	0.6996	0.7023	0.6985	0.7004	0.7013	0.7032
GTMP	0.3801	0.3787	0.3809	0.3798	0.3795	0.3787
GTMR	0.4021	0.4008	0.4029	0.4003	0.4017	0.3999
GTMF	0.3908	0.3895	0.3916	0.3898	0.3903	0.3890
DE						
SBLEU*	0.3600	0.3563	0.3604	0.3593	0.3566	0.3556
TER	0.8086	0.8108	0.8079	0.8084	0.8104	0.8109
GTMP	0.3217	0.3210	0.3218	0.3212	0.3210	0.3201
GTMR	0.3609	0.3602	0.3608	0.3590	0.3601	0.3580
GTMF	0.3402	0.3395	0.3402	0.3390	0.3394	0.3380

Table 2: Overall scores calculated using one set of reference translations, where * indicates averaged SBLEU scores.

Due to the low number of changes in the source, overall differences using automatic metrics are marginal. However, the scores obtained by the output corresponding to the grammar changes are consistently better than those obtained by the original set. A similar trend can be observed when comparing the Spelling scores with the Spelling+Grammar scores. While the Spelling set obtains worse scores than the Original set, the Spelling+Grammar set obtains better scores than the Spelling set (while being lower than the Grammar scores). These results suggest that the precision of the grammar suggestions is higher than that of the spelling and style suggestions. In order to find out how individual rules are behaving, a more granular scoring method is introduced in Section 4.2.

4.2 Individual Results for English to German translation

For the language pair English-German, Figure 2 shows how Acrolinx rules affect the scores of the automatic and human metrics. The number in parentheses indicates the number of reformulations. The metric LM-EN identifies the source language model rating, LM-DE identifies the target language model rating, AVG is the average of the BLEU, GTM and TER ratings, and HUMAN gives the human rating. (Here a rating of a metric specifies how many reformulations suggested have a better, worse, or equal effect on scores of that metric.)

It appears that human judgment seems to correlate with the automatic scores, except for language model scores. In fact, human perception gave more positive ratings on average to corrected translations than automatic scores. This result is not surprising, given the fact that the human score measures the quality of a translation against an entire language, instead of just a single reference sentence. Nevertheless, the correlation means that it seems reasonable to identify the suitability of a rule for pre-editing by the automatic scores alone, except for language model scores.

As for the language model scores, we could not see a helpful indicator in them. In general, the scores of both the source and target language rank reformulations much better than the other automatic scores and human perception. Even pre-editing rules that clearly have a bad or mediocre effect (such as *incorrect_extra_comma*) had an exceptionally good effect on the language model scores in some instances. For the target language, at least, an explanation is that the language model score only measures the fluency of the translation, but not the adequacy. Looking further into the cause is a topic of future work; we will not analyze these scores for now.

Note that the autoApplyClient currently does not distinguish between different spelling corrections. The input data set caused several hundreds of spelling corrections, which could not be feasibly evaluated by humans. As there was no distinction between the spelling corrections, the results would not have been representative anyway.

When looking at the type of error, there are some differences between automatic scores and human judgment. For each grammar and terminology rule, the average number of reformulations deemed “worse” is remarkably close to the number of reformulations regarded as worse by the human evaluator. For the remaining reformulations, the “better” and “equal” ratings differed between human and automatic scores, but not by much. For style rules, the human judgment gave more “better” and less “worse” ratings, but when looking at some rules in detail, there were some exceptions to this observation.

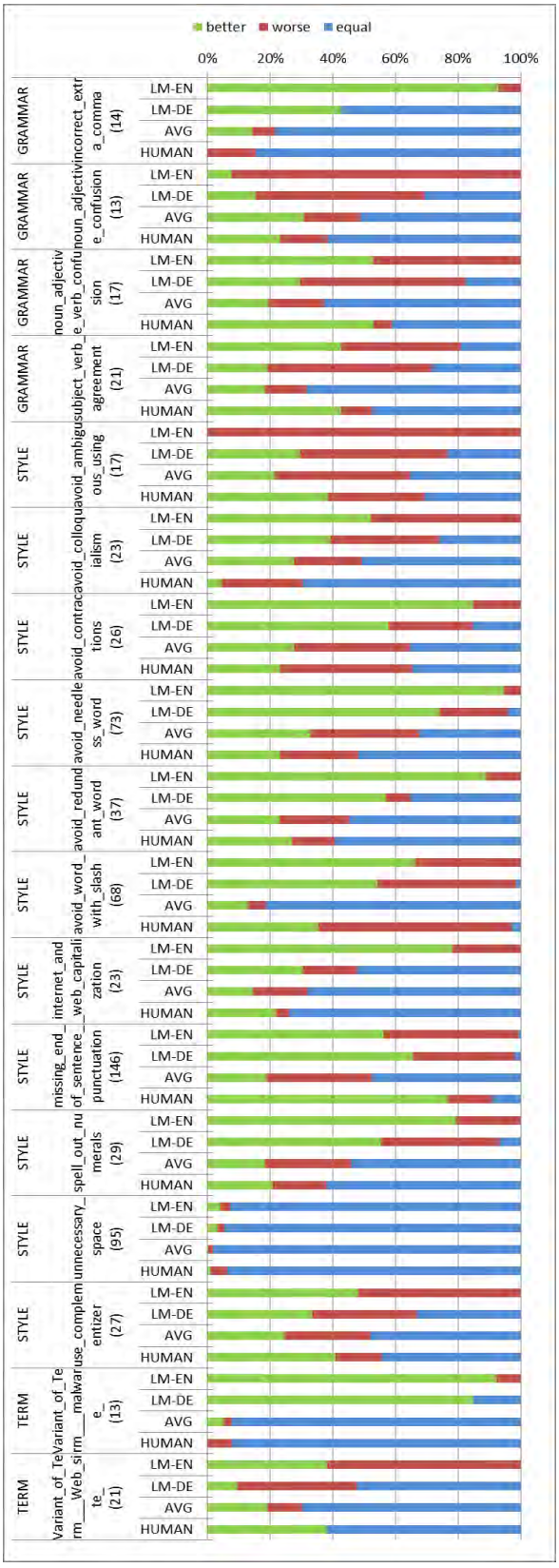

Figure 2: Individual German Results

For the rule *avoid_needless_word*, many translations automatically ranked as "better" have been judged "equal" by the evaluator. The reason is that the rule removes unnecessary filler words from a sentence, which are often part of colloquial language. Such filler words do get translated correctly, leading to an "equal" rating by the evaluator. However, colloquial language often allows for a range of possible translations, such that the probability of missing the reference translation is high. Due to the definition of the metrics, not including a word in the translation gets a better score than translating it "wrong".

The rule *avoid_word_with_slash* replaces the slash in sequences like "from/to" with a conjunction. This change was often considered better by the evaluator, even though it did not improve the automatic scores. At the same time, the rule also got an exceptionally high number of "worse" ratings by human judgment. It showed that the rule over-generates suggestions and also replaces slashes in menu descriptions like "Help / About" or directory names like "Windows/Temp".

Finally, the evaluator regarded the reformulations of the rule *avoid_colloquialism* more often "equal" at the expense of "better". In fact, the good ratings by the automatic metrics were often coincidence; the rule actually reformulated sentences in a way that did not improve the quality. Fortunately, it mostly also did not harm the translation quality.

Apart from these exceptions, we observe that the rule scoring framework is in general a reasonable approach to identify pre-editing rules by low "worse" values. Although a reformulation showed with a good "better" rating by automatic metrics sometimes had more "equal" ratings by humans instead, this effect was found to be rather small.

4.3 Individual Results for English to French Translation

Figure 3 shows the results for the translation of the segments from English to French. These results confirm the trend observed with English-German, whereby human judgment seems to correlate well with the average of the scores generated by automatic metrics, especially for the following rules: *unnecessary_space*, *incorrect_extra_comma*, *avoid_colloquialism*, *use_complementizer*, *avoid_contractions*, *avoid_redundant_word*,

avoid_ambiguous_using, *subject_verb_agreement*, and *noun_adjective_verb_confusion*.

However, some rules show inconsistent results between human judgments and automatic metrics. For example, the automatic scores obtained for the rule *avoid_needless_word* suggest an even distribution of improvements and degradations, whereas the human judgments suggest that degradations are much more frequent. Some of these *needless* words include adverbs such as "very" or "completely" which, when removed, affect the original meaning of the source text. This is particularly relevant in the context of User Generated Content since users will often rely on such words to express their level of frustration or satisfaction with a problem or solution.

Another rule which reveals inconsistent results is the *avoid_word_with_slash* rule, whose reformulations introduced problems when slashes are present in lists or product names (e.g. Norton Internet Security/Norton 360) or GUI options (e.g. "Exclusions/Low Risks Section"). In these cases, a reformulation with "or" tends to change the meaning of the original text. Perhaps more surprisingly, degradations were also introduced in running text (e.g. "Should I do a complete uninstall/reinstall?"). Regardless of the reformulation, either with "and" or "or", ambiguity problems were introduced and created translation problems.

Finally consistency differences emerge for rules whose reformulations introduce possible stylistic choices, such as the *spell_out_numerals* rule. For this rule, the number of improvements obtained with human judgments was lower than those obtained with automatic metrics. In some cases, changing a numeral (e.g. "3" in "for 3 weeks") with its full form ("three") introduced grammatical problems (such as a mistranslation of the preposition "for") while producing a semantically equivalent rendering of the numeral. In this particular case, this side effect was not captured by the automatic metrics.

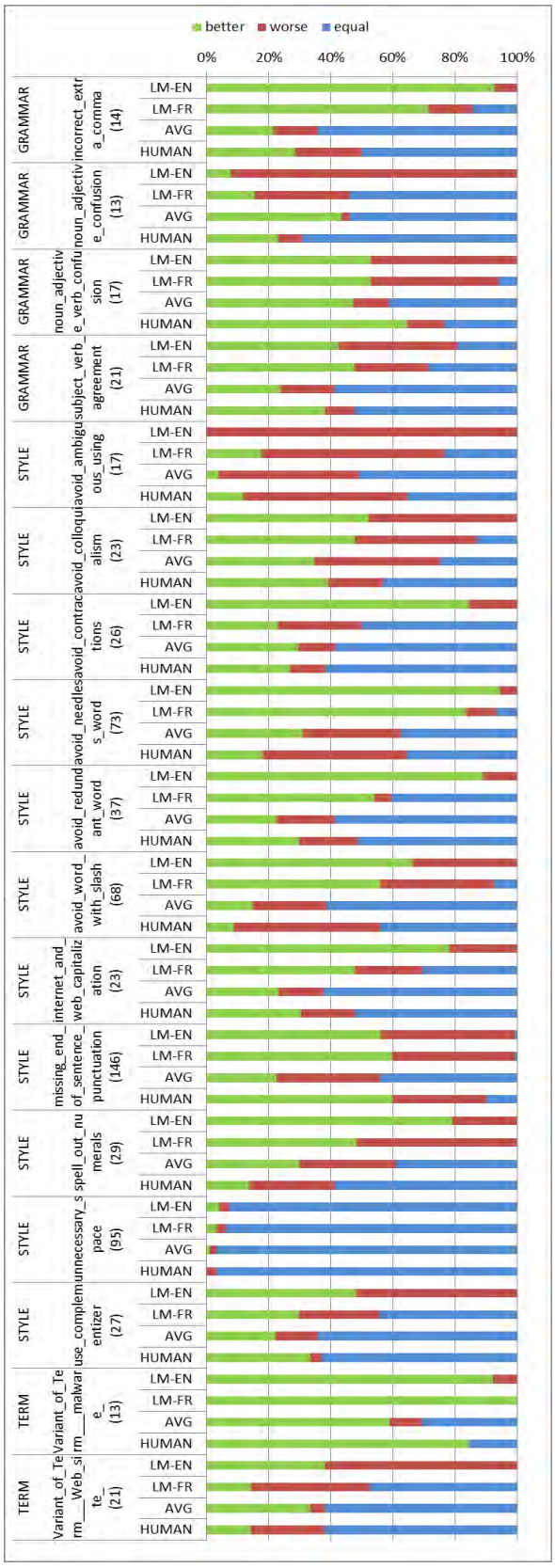

Figure 3: Individual French Results

5 Analysis

In this section we focus on an analysis of the rules that overall have the most positive impact (most improvements, less degradations).

5.1 English-German

To show that the framework is indeed suitable to identify beneficial or detrimental pre-editing rules, we looked at reformulations of rules that have notable ratings by automatic scores, (the AVG row of Figure 2). We looked at five rules, whose individual automatic score ratings are summarized in Figure 4.

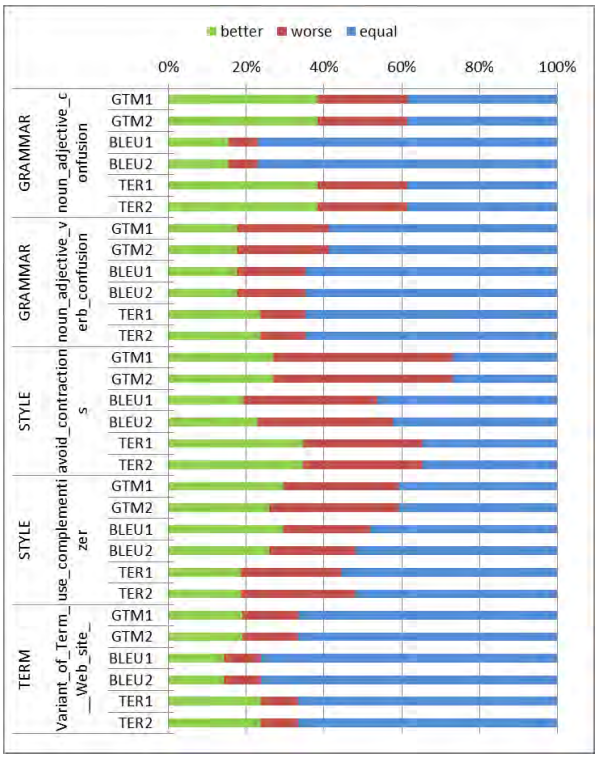

Figure 4: Most Effective Rules for German

Note that for each rule, the ratings of the individual metrics are relatively similar. It is thus a reasonable simplification to just examine the average, as we did in the previous section.

The rules *noun_adjective_verb_confusion* (17 occurrences) and *noun_adjective_confusion* (13 occurrences) both showed a very good result in the automatic scores. The rule changed the splitting of words according to the part of speech they form in the sentence, for example "a back up" is reformulated to "a backup", "the add on" to "the add-on", and "2 year contract" to "2-year contract". Usually, this helped Moses to pick the correct translation from the phrase table. The cases with a worse effect usually involved splitting a verb, like "please cleanup" to "please clean up", in sentences that require Moses to generate a long-distance dependency between the German verb and its prefix. Such discontinuous surfaces are challenging for non-hierarchical SMT translation systems, such as the one treated here. The surprisingly low BLEU ratings for *noun_adjective_confusion* is a coincidence.

The rule *use_complementizer* (27 occurrences) replaces phrases such as "make sure it works" with "make sure that it works". An included "that" obviously helped Moses to pick the correct segmentation of phrases.

The term variant of "Web site" was "Website" (21 occurrences), and this single word indeed helped Moses to choose the correct German translation "Website", and reduced the probability of "Web" and "site" being translated individually.

The results also clearly show that the rule *unnecessary_space* has almost no effect on the translation quality. This is not surprising: the rule removes mostly removes spaces that are removed by the tokenizer of the translation system anyway, so this pre-editing operation does not change the form in which sentences actually reach Moses.

The evaluation also shows that the rule *avoid_contractions* (26 occurrences) should probably not be used, because it has a mostly negative effect. On further examination, it showed that the rule often replaced "can't" with "cannot". In German translations of "cannot", Moses often left away the negation. Missing negations are particularly hard to detect by automatic evaluation metrics; they have a much worse impact on the translation than a missing article, for example, but the two cases cannot be distinguished by just examining the sentence surface. Therefore, it is an encouraging result that the evaluation actually hinted at problems with this rule.

Apart from the exceptions mentioned in section 4, where the automatic scores did not adequately reflect the translation quality as perceived by humans, the automatic framework thus helps to focus the manual evaluation on rules with a noteworthy effect.

5.2 English-French

The English-French results confirm what was observed for English-German, with an overlap of four rules, as shown in Figure 5.

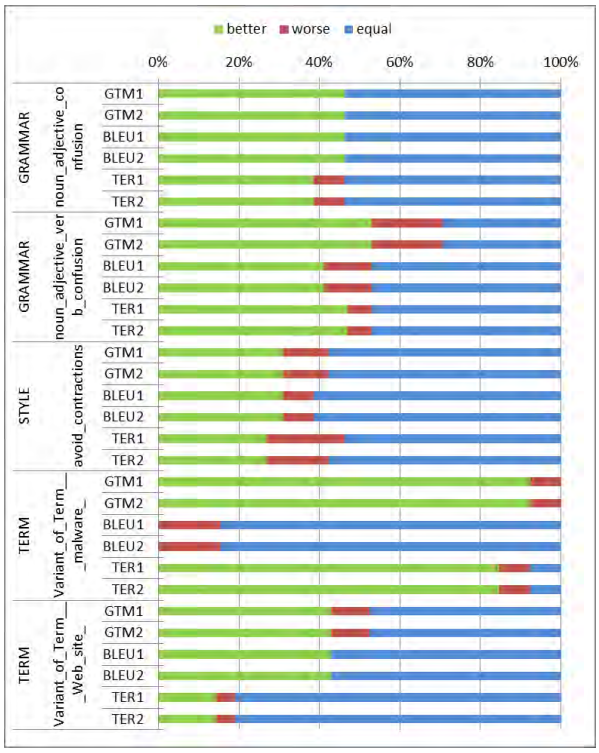

Figure 5: Most Effective Rules for French

Once again this framework proves useful, not only to identify rules whose reformulations generally improve translation quality, but also to identify cases where degradations are introduced. For instance, we were able to identify that the *avoid_contractions* rule introduced a problem when the contracted form "won't" (used to refer to a timeless state, as in "I have Firefox 3.1 and it won't work") is rephrased as "will not". In this case the resulting translation incorrectly conveys the notion of "future" ("fonctionnera" instead of "fonctionne").

Another example concerns the *noun_adjective_verb_confusion* rule, which overall improves translation quality. In a few cases, however, automatic scores are consistently degraded, especially when the rule triggers on words that are product names or feature names (e.g. "By the way, Cleanup sometimes requires a reboot to complete."). The reformulation of "Cleanup" as "clean up" generates a mistranslation, but this diagnostic helps refine the rule (either by adding an exception or by preventing the rule to trigger when words are capitalized in the middle of a sentence).

Finally, it is worth pointing out that while the various automatic scores provide consistent results for the grammar and style rules, differences emerge for terminological rules, such as the *malware_term_variant* rule (13 occurrences). This is due to the fact that BLEU appears far less sensitive than TER and GTM to the introduction of spurious words. In one example, changing "MALWARE" to "malware" resulted in two words being removed from the MT output ("infectés par"). While TER and GTM captured these changes by assigning different scores to the original and changed translations, the BLEU scores did not change. This confirms some of the inconsistencies observed with the BLEU Scores for English-German in Section 5.1.

6 Conclusions and Future Work

In this paper we have introduced a novel framework to analyze in detail the impact of source reformulations on machine translation quality. Rather than trying to improve an overall score produced by a given automatic metric, our main objective was to find out whether the development of such formulations could be improved and sped up using automatic metrics. Our results show that the scores generated by automatic metrics can be very useful to develop rules in general since their results are overall consistent with those provided by human evaluators. This is due to the fact that the reformulations introduced by these rules do not introduce changes which would create widely different translations (and thus contradict the use of a reference translation based on the original text). Using this framework, we have been able to identify rules that have an overlapping effect (either positive or negative) on the English-German and English-French language pairs. This is extremely useful in the context of User Generated Content because users of such rules are likely to expect high precision rules. Finally we have also been able to identify cases where rule refinements are required to further improve the precision of specific rules.

Besides refining existing rules based on this paper's findings, our future work will include further analysis to find out why the source and target lan-

guage model scores were unreliable indicators for translation quality and whether such scores could be useful to select a specific reformulation when multiple reformulations are available. We are also interested in investigating further whether scores generated by humans and automatic metrics correlate at the level of individual reformulations. Moreover, we would like to explore the use of lattice inputs instead of using individual reformulations, especially for those rules whose precision is not optimal. Looking further, another interesting line of work is to apply pre-editing rules to the source language side of the training corpus directly, to retrain the translation system, and to re-evaluate the impact of individual Acrolinx rules. Finally we would also like to extend this evaluation to other MT systems (including rules-based machine-translation systems) to find out to what extent pre-editing rules are system-independent.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 288769.

References

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

Bernth, A., & Gdaniec, C. (2002). MTranslatability. Machine Translation, 16, 175-218.

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., & Van Genabith, J. (2012). Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data? Proceedings of EAMT 2012. Trento, Italy.

Bredenkamp, A., Crysmann, B., & Petrea, M. (2000). Looking for Errors : A Declarative Formalism for Resource-Adaptive Language Checking. Proceedings of LREC 2000. Athens, Greece.

Collins, M., Koehn, P., & Kučerová, I. (2005). Clause restructuring for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05 (pp. 531-540). Morristown, NJ, USA: Association for Computational Linguistics.

Doherty, S. (2012). Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-methods approach. Dublin City University.

Genzel, D. (2010). Automatically learning source-side reordering rules for large scale machine translation. Proceedings of the 23rd International Conference on Computational Linguistics (pp. 376–384). Association for Computational Linguistics.

Leblanc, T. R. (2005). Is There A Translator In Teh House?": Cultural And Discourse Analysis Of A Virtual Speech Community On An Internet Message Board. Louisiana State University.

Lin, C.-Y., & Och, F. J. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain.

Melamed, I., Green, R., & Turian, J. (2003). Precision and recall of machine translation. Proceedings of HLT-NAACL 2003: Short Papers.

O'Brien, S., & Roturier, J. (2007). How Portable are Controlled Languages Rules ? A Comparison of Two Empirical MT Studies. Proceedings of MT Summit XI (pp. 345-352). Copenhagen, Denmark.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th Conference of the Association for Machine Translation of the Americas. Cambridge, Massachusetts.