Applying CNL Authoring Support to Improve Machine Translation of Forum Data

Sabine Lehmann, Ben Gottesman, Robert Grabowski, Mayo Kudo, Siu Kei Pepe Lo, Melanie Siegel, Frederik Fouvry

Acrolinx GmbH, Friedrichstr. 100, 10117 Berlin, Germany

Keywords: Controlled Natural Language, Authoring Support, Machine Translation, User-Generated Content, Forum Data

Abstract

Machine translation (MT) is most often used for texts of publishable quality. However, there is increasing interest in providing translations of user-generated content in customer forums. This paper describes research towards addressing this challenge by automatically improving the quality of community forum data to improve MT results.

1 Introduction

With the exception of some experiments such as Roturier (2006), authoring support and machine translation (MT) have generally been seen as two distinct areas. We want to show that MT can profit from being combined with authoring support methods.

Authoring support with language technology methods is aimed at supporting authors in the process of writing controlled natural language (CNL). Kohl (2008) gives guidelines for writing text that is easily translatable – by humans as well as by machines. Thicke (2011) measures the effect on MT quality of improving the source-language text by applying the recommendations of the Global English Style Guide. Tools based on computational linguistic methods, such as Acrolinx (www.acrolinx.com), implement these rules and thus support technical documentation writers. These tools provide spell and grammar checking, style checking, term extraction and terminology checking, and sentence clustering.

Many users of authoring support software are already using it in conjunction with translation memory tools such as TRADOS (www.trados.com) (Somers, 2003). The translation memory tools provide access to previously translated sentences, based on fuzzy matching algorithms.

adfa, p. 1, 2011.

These users are gradually adopting MT software as well. Often, however, they are not aware of the possibilities and limitations of the different MT methods and tools, and this leads some to reject MT after a period of experimentation.

Thus, some users already combine the tools in their daily work, but the tools themselves are completely distinct.

In the domain of technical authoring, pre-editing has previously been used to improve the human- and machine-translatability of the source text. Studies such as O'Brien and Roturier (2007) and Aikawa et al (2007) have shown that this approach can improve machine translation quality.

The situation in our case is different from the technical authoring situation: The text that we want to translate is user-generated content (UGC) from customer forums and is therefore written in a language quite different from (relatively standardized) technical documentation. Furthermore, it is impossible to train a statistical machine translation (SMT) system on this type of data because there is not enough human-translated forum text available as training data. As a result, we are forced to translate UGC using a mismatched SMT system trained on technical documentation. One goal of pre-editing is therefore to make the source language text more similar to technical documentation text.

Experiments (such as Allen (2001) and Plitt and Masselot (2010)) show that manual post-editing on high-quality MT output is already much faster than translating from scratch by humans. We aim to reduce the manual post-editing effort even further through the use of automatic rules that support the post-editing process and by enhancing the quality of the MT output.

Researchers have become interested in a variety of aspects of post-editing, such as the cognitive effort it takes (Temnikova 2010), productivity gains (Garcia 2011; de Sousa et al 2011), or how to best support human workflows by translation quality estimation (Specia 2011).

Few approaches to automatic post-editing have been discussed in the literature, among them: using statistical methods (Dugast et al 2007; Simard et al 2007) or manually built regular expressions (Guzmán 2008). All approaches are first steps in a developing field of research.

The idea of using post-editing patterns as error descriptions appears in Stymne and Ahrenberg (2010). The authors use a grammar checker both to analyze SMT output and to automatically correct it. Although the reported experiments serve as a proof of concept, the authors note that the grammar checker they used was built for supporting human document production and thus was unable to spot many errors typical of MT systems.

2 Background

2.1 Authoring Support Software

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. It is the most commonly used software in the technical documentation authoring process, where reformulation suggestions are applied by professional writers.

2.2 Methods in Machine Translation

We look at two approaches to machine translation: the statistical approach and the rule-based approach. Although there has been much research on combining the ideas of both, such as Eisele (2009), we make a clear distinction at this point so that it is easier to evaluate the influence that authoring support has on each.

The rule-based approach to MT makes use of linguistic information and dictionaries and defines translation rules based on this information. Examples of systems using this approach are Lucy (www.lucysoftware.com), Langenscheidt T1 (www.langenscheidt.de), and Systran (www.systransoft.com). We expect a positive effect on the MT output when source language texts are more correct and therefore easier to analyze.

The statistical approach to MT (Koehn, 2010) analyzes large amounts of parallel data. On the basis of this training data, statistical models are built, such as the 'phrase table', which consists of probabilities of given source-language phrases being translated as given target-language phrases. These models are then used in the translation task. One source of such parallel data is translation memories. Examples of systems implementing this approach are Google Translate (translate.google.de) and Moses (www.statmt.org/moses). We expect a positive effect on the MT output when the input is standardized, such that it diverges less from the training data.

In machine translation research, evaluation often consists of comparison of machine translation output with a 'gold standard' reference translation, with the assumption that greater similarity is better. Some of the similarity metrics standardly used for this are Smoothed BLEU (Bilingual Evaluation Understudy) as described by Lin and Och (2004), GTM (General Text Matcher) f-measure as described by Melamed et al (2003), and TER (Translation Error Rate) as described by Snover et al (2006). We use these scores both to identify sentences which are difficult to translate, so that requirements for reformulations can be implemented in pre-editing rules in a more focused and targeted way, and to evaluate the effectiveness of pre-editing and post-editing.

2.3 User-Generated Content

Software publishers rely on manuals and online support (knowledge base) articles to assist their customers or 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. These conversations take place in a number of online environments (e.g. stackexchange.com), some of which are moderated and

facilitated by software publishers. 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, major 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 of the language used in the forum (so its linguistic and technical accuracy may not be optimal).
- Although written, this content is stylistically closer to spoken language, with informal syntax and creative lexicon.
- Some of the content is authored by power users (or "techies") who "exhibit communicative techniques and practices that are guided by attitudes of technological elitism" (Leblanc, 2005). These can include alternative spellings, acronyms, case change, color change, techie terms, emoticons, or representation of non-lexical speech sounds.

3 The Problem

Due to the characteristics of user-generated content described in the previous section, translation systems usually do not translate it well. We aim to show that this problem can be ameliorated using Acrolinx. The latter can be used both to reformulate source language forum entries in order to make them more machine-translatable (pre-editing) and to reformulate the translated entries in order to correct translation errors. Where possible, these reformulations should be done automatically. For cases where an automatic reformulation is not possible, we also investigate ways to support forum users via authoring support mechanisms. Here, the major challenge is to adjust the typical CNL workflow to average forum users, who tend to be non-professional writers.

However, authoring support software, like machine translation software, is not built for dealing with UGC, nor is it built for machine translation output. While the standard "out-of-the-box" Acrolinx checking rules help the author to improve the quality of the content in the source language, they need to be specially adapted to also have a positive effect on the machine-translatability of the text. Likewise, a rule set for post-editing machine translation output needs to be specially designed. The main task is to identify which rules to choose, and how to adapt them.

4 Methodology

4.1 Pre-Editing

We conducted a standard corpus analysis of the forum text to identify areas where source normalization is required to ensure that input text is as understandable and consistent as possible and matches as much as possible the coverage of the translation system. Such an analysis covers the following aspects:

Spelling and grammar

Using authoring support in the pre-editing process firstly involves correcting spelling and grammar errors in the source document. We expect that correcting spelling and frequent grammar errors such as agreement errors will improve the translatability of the source text. Automatic spelling correction is, however, error-prone, as there is usually more than one correction suggestion for a misspelled word.

Terminology

Consistent and precise use of terminology also helps the MT process. Acrolinx provides functionality both for extracting terminology in order to set up a terminology database, and for checking terminology usage against such a database. Term extraction rules are based on linguistic information and run on data in the relevant domain. Thus, the extracted terms are more useful than, for example, a general domain-independent ontology.

In forum data, authors tend to use terminological variants, such as short names for products, that do not occur in the more formal sorts of text that are used as SMT training data. Therefore, it is important to extract terminology from forum data before checking terminology in forum data.

Style

Authoring support for technical documentation includes rules regarding style of writing. These rules are designed to improve the consistency, clarity, understandability, and (human-)translatability of text. Many of these rules also have a positive effect of machine-translatability. For example, rules that simplify long sentences and complex syntactic structures lower the burden on the MT system. Furthermore, we developed rules specifically aimed at improving the machine-translatability of forum text, such as *avoid_colloquial_language*.

An outcome of our analysis was that we identified language peculiarities of forum data that need to be reflected in the basic linguistic resources, such as specific words or types of errors. Many of these are due to the informal or spoken-language style used. Here are some examples from the English and French forums:

- It was **some what** messy.

- Re: symantec update **wont** work.
- It's very interesting, **wanna** give it a go?
- 512MO ram de **dique** dur, mais **la**, cela a toujours **fonctionner** normalement avant, cela fait 4 jours que le **probleme** est apparu quand des mises **a** jours Windows ont été faites.

French forum writers often commit the error of omitting an accent. There are multiple instances of this in the last example above. Often, this is seen by authoring support software as a mere spelling error, as in the case of "probleme" vs. "problème". In other cases, however, the author in effect substitutes a homophone for the intended word. One of the classic homophone pairs is "à" and "a", the first of the two being a preposition and the second a form of the verb "avoir" ("has"). Given that there are a series of such "confusion words", a lot of the work on French has focused on developing rules to correct this sort of error.
Using this information, we were able to identify and implement a first set of pre-editing rules for English and French forum data, such as:

- (FR) Confusion de mots (la vs. là, ce vs. se, etc.) (word confusion)
- (FR) Mots simples (simple words)
- (FR) Évitez le langage familier (avoid informal language)
- (EN) do not use complex sentences
- (EN) do not omit relative pronouns such as that and which
- (EN) use standard capitalization
- (EN) avoid parenthetical expressions in the middle of a sentence

The identification of those rules was based on extensive manual and semi-automated data analysis. The work was carried out in several cycles and has focused on one first set of data. The next step will focus on a different corpus in order to verify whether the basic findings can be confirmed.

The pre-editing rules can on the one hand be applied by authors, as is usually done in the technical documentation authoring process. The author gets error flags and decides about reformulations. This process ensures that text reformulations are always correct. Further, a learning process for the author starts. He or she gets a better understanding of the abilities and limits of MT. On the other hand, many of these rules can be automatically applied, which saves a lot of time and human effort. It lowers the precision, but this may be an acceptable tradeoff as long as the overall effect on the translation quality is positive.

4.2 Post-Editing

The authoring support tool is applicable to monolingual text. Therefore, using the same mechanisms as in pre-editing, we can correct spelling, grammar and style errors on the target language text. In order to identify errors to correct via automatic post-editing, we conducted experiments involving professional translators performing post-editing on MT output. Aside from standard spelling and grammar corrections, we

identified rules for automatic post-editing of German machine translation output on this basis. Here are examples of these rules:

- correct terminology
- correct standard expressions
- correct word order
- convert future tense to present tense
- convert indicative to causative
- convert "man" to passive
- convert series of "von"+ Noun

Terminology errors account for a large percentage of the corrected errors and thus deserve special attention. Allen (2001) describes a tool that supports humans in manually adding unknown words to a dictionary and in domain-adaptation by manually selecting the best translation of words in a certain domain. We propose to do multilingual term extraction using Acrolinx, and make use of term bank organization (with domains and linked terms) for domain-dependent selection of translations. Domain-specific terminology can be extracted from the training data, a term bank is set up with information about the domain and used for terminology checking on the source and target language texts.

4.3 Support with Automatic Tools

For many errors found in a source text, there is an obvious improvement suggestion. In fact, many of the developed Acrolinx rules for pre-editing suggest the appropriate improvement to support the author. It is thus feasible to include automatic reformulations in the translation workflow. The challenge is that there may be several possible improvement suggestions for a detected error (a common situation for spelling mistakes in particular), or there may even be several overlapping error flags that have to be resolved in a particular order.

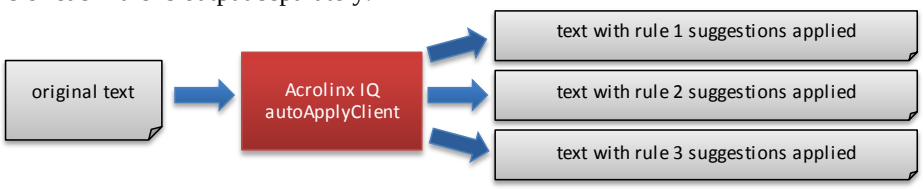
In order to perform automatic pre-editing tasks, and to evaluate their impact on translation quality, two automatic tools have been developed.

AutoApply client

The Acrolinx AutoApply client is an Acrolinx client that checks a document and automatically replaces flagged (marked) sections of text with the top-ranked improvement suggestion given by Acrolinx, provided that the flag has at least one such suggestion. The tool can be restricted to applying only the suggestions of specific error types, rule sets, and term sets to a given document, such that their respective impacts can be examined in isolation.

Rule scoring framework

To fully automatically identify the suitability of rules and suggestions for pre-editing purposes, we have developed an automatic rule scoring framework. It first uses the AutoApply client to automatically apply the suggestions of rules to the sentences to get reformulated versions of the sentences. The result of applying suggestions of each rule is output separately.



We then machine-translate the reformulated and corresponding original sentences, and score the translated sentences against the corresponding reference translations using the metrics mentioned in section 2.2. Since each reformulation reflects the application of only one suggestion, the differences in the scores are usually rather small. Therefore, we merely perform an automatic contrastive evaluation: we note whether the scores have improved, degraded or stayed the same after the reformulation according to the automatic metrics. The mechanism can be complemented by a human evaluation, where judges are given the task of judging which of the pair of translations pair is superior. The rating results are grouped by rule name, such that the rules with the best impact on the translation quality can be easily identified. This automatic analysis thus focuses the manual task of finding new potential rule candidates on "interesting" rules.

5 Evaluation

The following table summarizes the results for some of the automatically applied Acrolinx rules for translations from English to French. The results are relative to the instances in which a rule could be applied. The table indicates for each scoring metric in how many of these instances the reformulation of the source segment resulted in a better translation ("+" column), and in how many instances it resulted in a worse translation ("-" column). Note that for the remaining instances, the translation quality did not change.

	Human evaluation		GTM		BLEU		TER	
	+	-	+	-	+	-	+	-
Use relative pronouns such as that and which	33 %	4 %	26 %	19 %	26 %	7 %	15 %	15 %
Confusion between noun and adjective	23 %	8 %	46 %	0 %	46 %	0 %	38 %	8 %
Avoid contractions	27 %	12 %	31 %	12 %	31 %	8 %	27 %	19 %
Internet and Web capitalization	30 %	17 %	30 %	22 %	22 %	9 %	22 %	13 %

Two conclusions can be drawn from the table. First, the four rules shown have a predominantly positive impact on the translation quality, and can thus be used in a fully automatic pre-editing mechanism. Second, and more importantly, the ratings based on the calculated automatic scores correlate well with human judgements. This leaves us confident that we can use the automatic framework to easily identify suitable pre-editing rules, and possibly to support the development of rules that have a positive impact on translation quality.

6 Summary

We have shown how CNL authoring tools can support and supplement the process of machine translation of user-generated content. The ACCEPT research project aims at translating user-generated content automatically with a statistical MT system that is trained on user manuals. This divergence of training data and test data is due to the fact that there simply is not enough human-translated UGC available to train an SMT system. Pre-editing rules implemented in the CNL authoring tool are aimed at reformulating the source text to make it easier to translate automatically. Therefore, we first implemented corrections of spelling and grammar, specifically adapted to the data. When using statistical MT, the translation quality is improved if corrections to the input text make it more similar to the training corpus. In a next step, we implemented pre-editing rules that stylistically reformulate the source text, such as by shortening sentences. We have shown a procedure to analyze the corpus, identify pre-editing rules, implement them and evaluate their impact.

Post-editing rules are applied to the MT output. They reformulate the output text to fix common mistakes by the MT system. These rules concern grammar, spelling, and standard stylistic reformulations as well as terminology corrections.

The automatic application of CNL reformulations seems feasible and can be part of the machine translation process. We implemented an automatic tool to support the identification of suitable pre-editing rules that can be safely automatically applied, and that clearly improve the translation quality. This tool is also used to support the identification and development of new rules, even for manual pre-editing workflows.

The next step will be to conduct more evaluations and therefore get a better insight into the impact of specific rules on the MT process. Further, we will evaluate the translation of more language pairs.

Future work will focus on analysis of the training corpus. The idea is to investigate whether the training corpus provides clues about which rules would be best to use. More concretely, it would be a sort of "reverse engineering": if rules don't trigger in the training corpus or trigger very rarely, then it is likely that the structure or pattern the rule is looking for does not occur in the training corpus. In that case, we would assume that this rule might be a good rule candidate for pre-editing. On the other hand, if a rule triggers a lot in the training corpus, it means that the structure or pattern in question is frequent in the training corpus. As a consequence, we would assume that the MT engine has no problem translating them and we would not take it up as rule candidate. So we hope that the trigger frequency correlates with its usefulness in pre-editing.

The advantage of such an approach is that it would help us automatically select different rules depending on which training corpus we use – a feature which is highly valuable as it would allow us to take into account the specificities of different MT systems.

Acknowledgments

This project is carried out with financial support from the European Community through the 7th Framework Programme for Research and Technological Development (grant agreement 288769).

7 Bibliography

T. Aikawa, L. Schwartz, R. King, M. Corston-Oliver, & C. Lozano, (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.

Jeff Allen (2001): Postediting: an integrated part of a translation software program. In: Language International, April 2001, pp 26-29.

Sheila de Sousa, Wilker Aziz, and Lucia Specia (2011): Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles. In Proceedings of the International Conference Recent Advances in Natural Language Pro-

cessing 2011, pages 97–103, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Loic Dugast, Jean Senellart, and Philipp Koehn (2007): Statistical post-editing on SYSTRAN's rule-based translation system. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 220–223, Prague, Czech Republic, June. Association for Computational Linguistics.

Eisele, A. (2009): Hybrid Architectures for Better Machine Translation. GSCL WS „Kosten und Nutzen von MT", Potsdam, September 2009.

Ignacio Garcia (2011): Translating by post-editing: is it the way forward? Machine Translation, 25:217–237. 10.1007/s10590-011-9115-8.

Rafael Guzmán (2008): Advanced automatic mt postediting. Multiling Computing, 19(3):52–57, April.

Philipp Koehn (2010): Statistical Machine Translation. Cambridge University Press.

John Kohl (2008): "The Global English Style Guide. Writing Clear, Translatable Documentation for a Global Market". Cary NC: SAS Institute INC.

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.

M. Plitt, F. Masselot (2010): A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. In The Prague Bulletin of Mathematical Linguistics 93, pages 7-16.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn (2007): Rule-Based Translation with Statistical Phrase-Based Post-Editing. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 203–206, Prague, Czech Republic, June. Association for Computational Linguistics.

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.

H. L. Somers (2003). Computers and translation: a translator's guide, chapter Translation Memory Systems, pages 31–48. John Benjamins Publishing Company, Amsterdam, Netherlands.

Lucia Specia (2011): Exploiting objective annotations for measuring translation post-editing effort. In Proceedings of the 15th International Conference of the European Association for Machine Translation, pages 73–80, Leuven, Belgium.

Sara Stymne and Lars Ahrenberg (2010): Using a grammar checker for evaluation and postprocessing of statistical machine translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, may. European Language Resources Association (ELRA).

Irina Temnikova (2010): Cognitive evaluation approach for a controlled language post-editing experiment. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, may. European Language Resources Association (ELRA).

Lori Thicke (2011): Improving MT results: a study, (Multilingual 22(1):37–40).