Comparison of alternatives to strict source control: a case study with -ing words.

Nora Aranberri[1] and Dr. Johann Roturier[2]

[1] SALIS, Dublin City University, Glasnevin, Dublin 9, Ireland
nora.aranberrimonasterio@dcu.ie
[2] Symantec Ireland, Ballycoolin Business Park, Blanchardstown, Dublin 15, Ireland
johann_roturier@symantec.com

Abstract. This paper presents the results of a case study focusing on the design and development of specific Controlled Language rules, which address readability and translatability issues generated by -ing words. This paper reviews several development and deployment iterations, whereby users' feedback was taken into account for future development phases. We show that two complementary techniques may be used: the generation of flags for rule violations (and subsequent human validation) and the fully automatic introduction of additional elements in input text (pre-processing). We compare and contrast these two techniques in this paper, showing that the precision of both our machine-oriented rules and human-oriented rules reaches at least 80%.

Keywords: controlled language, machine-translatability, ambiguity, -ing, case-study

1 Introduction

The present study is conducted in the context of an implementation of a set of controlled language rules within the authoring process of software-related technical documentation. This documentation set, which may be published onto the Web, must also be translated in a number of languages using a rule-based machine-translation (RBMT) system. We focus on controlling -ing words for two main reasons: their high frequency (2.1% of words in a half million word corpus) and their reported ambiguity [1]. The motivation for controlling the use of -ing words is due to their flexibility. They can serve different grammatical functions while keeping the same form, so this poses problems of source language analysis for both human readers and machine translation systems. Kohl [2] identified three reasons why -ing words pose problems for readers: grammatical function flexibility, ungrammatical usage and ambiguity. In addition to the difficulties for human readers, multiple target structures may be specified for different -ing word classes/structures [3] thus creating translation difficulties. Also, the lack of semantic knowledge may prevent an RBMT system from assigning the correct syntactic information [4]. Besides, O'Brien [5] found that six out of the eight CLs she analysed shared a rule which recommended avoiding gerunds. With the aim of finding ways not to flag widely used grammatical structures, we investigate the possibilities of generating flags for rule violations (and subsequent human validation) and the fully automatic introduction of additional elements in input text (pre-processing).

We partnered with an information development team that specializes in the development of XML topics in order to turn specific style guidelines into controlled language rules for English. While these topics adhered to a subset of the DocBook XML DTD presented by Walsh and Mueller [6], the text used within specific XML elements did not conform to any formal rules. The absence of strict CL rules made it difficult for various stakeholders (such as editors, managers and localisation teams) to monitor the status of specific topics based on the adherence to these rules.

2 Background

This case study builds up on Roturier's work [7], which focused on the development and deployment of a rule that attempted to (1) remove the ambiguity created by specific -ing words to improve the readability of source XML topics and (2) avoid certain -ing words to improve the performance of an RBMT system. The rule was designed to maximize recall in order to give writers the chance to address as many problems in the time allowed. The main learning points were as follows:

1. Some of the writers did not possess advanced linguistic skills and it proved difficult to explain in plain terms what the rule was supposed to achieve, leading to confusion on occasions.
2. A lack of synchronization was noticed between the terminological exceptions that the controlled language rule required and terminology resources. The terminology repository was not updated sufficiently frequently for false positives to be eliminated efficiently.
3. The precision of the rule was not high enough for the rule to be used as a decision point in a workflow (such a decision point would indicate that if a topic contained less than X violations of the rule, this topic should not proceed to the next step). False positives sometimes remained in XML topics but it was not possible to mark as them "reviewed".

The present case study tried to refine the initial rule using a two-pronged approach. These two complementary approaches are described in sections 3 and 4.

3 Generating Controlled Language rule validation flags

The main advantage of CLs for translation is that by modifying the source, a number of target languages can benefit from it. However, if the error is only found in one of the target languages, or the source text does not incur into any grammatical or stylistic violation, the use of CL rules might not be the most appropriate. It might have an adverse effect in other target languages. Also, it will mean an increased effort from technical writers trying to comply with unnatural restrictions.

We performed a human evaluation to examine the machine translation of -ing words [8]. Results showed that 70-80% of the examples were correctly translated into French, German, Japanese and Spanish [9]. Within the incorrect translations, both language-specific issues and issues shared across the target languages were observed. Based on these findings, we propose to add three style rules to the acrolinx™ rule formalism [10], which takes a phenomena-oriented approach, that would mark ambiguous -ing words:

1. A rule that identifies reduced relative clauses, that is, the cases where a noun is post-modified by an -ing word, and suggests the use of a that/which clause. The pattern "noun + -ing word" is very common and therefore, we propose a lexical rule for the most recurrent -ing words in such structures. This results in a high precision (91.67%) but a lower recall (85.86%). An example is provided below:

Uncontrolled input: *The drive containing the Backup-to-disk folder is full.*
Controlled input: *The drive that contains the Backup-to-disk folder is full.*

2. A rule that identifies dangling subordinate subjects. It retrieves instances where the subject of the main and subordinate clauses are different and suggests making the subordinate subject explicit. Identifying the subject of an implicit structure is difficult. We therefore propose a broad rule which disregards second person singular implicit subjects and compromises precision in favour of recall (precision 84.61%, recall 92%).

Uncontrolled input: *These steps are not necessary when using the Backup Exec push installer.*
Controlled input: *These steps are not necessary when you use the Backup Exec push installer.*

3. A rule that identifies missing articles in the context of -ing words. Articles mark the beginning of noun phrases and are, therefore, effective markers for disambiguation between gerund-participles and participial adjectives. However, according to grammar not all noun phrases can have an article, which decreases the potential of this rule for disambiguation purposes (precision 80%, recall 85%).

Uncontrolled input: *Configuring notification in CASO*
Controlled input: *Configuring a notification in CASO*

4 Pre-processing

In section 3 we showed that only specific structures could be addressed by high precision rules when these rules are to be used by technical writers. To deal with language-specific problems, we use a pre-processing approach to transform input text in specific situations. Automatic pre-processing is a concept that has been described in [11] and [12]. The former explored parse tree reordering of the source while the latter approached source control as the machine translation from natural to corrected source. Our approach differs from these implementations since it is based on the following objectives:

- We want to annotate input text only in specific situations, e.g. when the source text should be machine-translated into Japanese in a particular context. To achieve this objective, we want to create a rule that will generate a reformulation of the original text. This reformulation may contain extra words, modified words, or extra syntactic annotation (such as specific POS tags).
- We also want to preserve the original text to make sure that the integrity of the source asset is not affected. For instance, we do not want to impact Translation Memory leverage at any stage of the translation workflow.

We present some of the technical details of the pre-processing approach we implemented. This implementation is based on the following sequence of steps:

1. First, we design a rule that generates a reformulation for each ambiguous or ungrammatical structure.
2. Before sending an input asset to our machine-translation system, we check whether the source asset contains any violations of the rules defined in step 1. If violations are found, an XML report containing reformulations will be generated.
3. The XML report is parsed to identify segments and their reformulations. These segments are matched against the assets of the original asset in order to produce a temporary source asset that contains original source segments and reformulated source segments.
4. The modified source asset is sent to our MT system that has been configured to analyse modified source segments as input text. The MT system will translate these segments and generate translated strings.
5. The original source segments are then paired with the machine-translated strings before being sent to the next step of a localisation workflow.

This approach allowed us to achieve the objectives defined earlier. We present the evaluation results of specific rules. We decided to address the handling of subordinate clauses introduced by “when + -ing” when translating from English into French as suggested in [8]. A pre-processing rule was developed to introduce a subject (“you”) and transform the gerund into the present form of the verb. We also decided to create a rule to identify gerunds (VBG) followed by noun phrases with plural, proper and uncountable head nouns. We found that we were able to add accurate syntactic annotation to these words 98.58% of the time, using an evaluation sample containing 300 segments.

The main advantages of this pre-processing approach are two-fold. First, these rules do not require any human validation so they do not impact on the productivity of writers. Second, the reformulations do not have to be completely grammatical. For instance, this approach allows introducing an article in front of a mass noun or a plural noun without impacting the acceptability of the source text. However, this approach also has limitations since it cannot address MT generation issues.

References

1. Bernth, A. and Gdaniec, C.: MTranslatability. *Machine Translation*. 16, pp. 175--218 (2001)
2. Kohl, J. R.: The Global English Style Guide: Writing Clear, Translatable Documentation for a Global Market. SAS Institute Inc., Cary, NC (2008)
3. Izquierdo, M.: Análisis contrastivo y traducción al español de la forma -ing verbal inglesa. Master Thesis. University of León, Spain (2006)
4. Arnold, D.J., Balkan, L., Meijer, S., Humphreys, R. L., Sadler, L.: Machine Translation: an Introductory Guide, Blackwells-NCC, London (1994)
5. O'Brien, S.: Controlling Controlled English. In: Proceedings of CLAW, pp. 105--114 (2003)
6. Walsh, N., Muellner, L.:DocBook: The Definitive Guide. O'Reilly & Associates, Inc. (2006)
7. Roturier, J.: Assessing a set of Controlled Language rules: Can they improve the performance of commercial Machine Translation systems? In: ASLIB Proceedings, pp. 1--14 (2004)
8. Aranberri, N., O'Brien, S.: Evaluating Machine Translation output for -ing Forms: a quest for improvements in four target languages. In: Evaluation of Translation Technology, Linguistica Antverpiensia (forthcoming)
9. Aranberri, N.: Exploding the Myth: the Gerund in Machine Translation. In: Proceeding of the LRC XII, pp. 211--219 (2007)
10. Bredenkamp, A., Crysmann, B., Petrea, M.: Looking for Resource-Adaptive Language Checking. In: Proceedings of LREC, pp. 667--673 (2000)
11. Adriaens, G.: Simplified English Grammar and Style Correction in an MT Framework: The LRE SECC Project. In: ASLIB Proceedings, pp. 73--82 (1994)
12. Collins, M., Koehn, P., Kucerova, I.: Clause re-structuring for statistical machine translation. In: Proceedings of the 43rd ACL, pp. 531--540 (2005)