MuTUAL: A Controlled Authoring Support System Enabling Contextual Machine Translation

Rei Miyata
The University of Tokyo
rei@p.u-tokyo.ac.jp

Anthony Hartley
Rikkyo University

Kyo Kageura
The University of Tokyo

Cécile Paris
Data61, CSIRO

Masao Utiyama
NICT

Eiichiro Sumita
NICT

Abstract

The paper introduces a web-based authoring support system, **MuTUAL**, which aims to help writers create multilingual texts. The highlighted feature of the system is that it enables machine translation (MT) to generate outputs appropriate to their functional context within the target document. Our system is operational online, implementing core mechanisms for document structuring and controlled writing. These include a topic template and a controlled language authoring assistant, linked to our statistical MT system.

1 Introduction

For improved machine translatability, a wide variety of controlled language (CL) rule sets have been proposed (Kittredge, 2003; Kuhn, 2014). Evidence of reduced post-editing costs when a CL is employed is provided (Bernth and Gdaniec, 2001; O'Brien and Roturier, 2007), and several controlled authoring support tools, such as Acrolinx[1] and MAXIT[2], have been developed. The fundamental limitation of the CLs proposed hitherto is, however, that they are defined at the level of the sentence rather than at the level of the document (Hartley and Paris, 2001). In fact, the notion of functional document element (see Section 2.1) does figure in some CL rule sets. ASD Simplified Technical English (ASD, 2013), for example, specifies writing patterns linked to functional roles of the document elements; the recommended maximum length of sentence is 20 words for 'procedural' writing and 25 words for 'descriptive' writing. Yet, the granularity of the elements is not high enough to enable detailed definitions of linguistic patterns within the elements. Thus it is necessary to formalise a document-level framework which enables context-dependent CL specification.

In this paper, we introduce an integrated web-based system, **MuTUAL**, which implements a suite of controlled authoring support modules, combined with our statistical machine translation (SMT) system. At the document level, document structuring modules help authors create well-organised documents. At the sentence level, controlled writing modules help them write source texts (ST) consistent with source-language CL rules. The principal innovation in the system is to contextualise the CL rules in the document structure to enable MT to generate outputs consistent with the target-side CL for a given functional element. While the current system supports the creation of municipal procedural documents in Japanese and their translation into English, it is extensible to other language pairs and text domains.

2 Contextual Translation

MuTUAL starts from the observation that the same source sentence should be translated as different target sentences depending on its location within the functional elements of the document. Let us consider this example Japanese sentence from a procedural technical manual: '文書を印刷する*/bunsho o insatsu suru*'. This sentence can appear as a task title in a section heading or as a step description in an itemisation, and should be translated, respectively, as 'To print a document' or 'Print the document'. That is, the translation depends on the item's functional role within the document.

[1] http://www.acrolinx.com/

[2] http://www.smartny.com/maxit.html/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 35–39, Osaka, Japan, December 11-17 2016.

DITA element in **Body** (default)		Specified functional element
Prereq	information the user needs to know before starting	**Personal condition** **Event condition** **Item condition**
Context	background information	**Explanation (Summary, Purpose, Expiration of validity, Penalty, Related concept)**
Steps	main content: a series of steps	**Necessary items to bring** **Place to go** **Form(s) to complete**
Result	expected outcome	**Result (Period for procedure, Items to be issued, Contact from local government)**
Postreq	steps to do after completion of current task	**Guidance to other procedures**

Table 1: Instantiation of the DITA Task topic

To realise contextual translation using MT, we (1) formulated a document structure for municipal procedures based on the Darwin Information Typing Architecture (DITA) framework (OASIS, 2010), and (2) defined context-dependent CL rules in both source and target languages according to the functional document elements, in combination with ST transformation rules.

2.1 Functional Task Elements

DITA is an XML architecture for authoring and publishing technical information which supports structured authoring to help writers compose a modularised chunk of information, called *topic* (Bellamy et al., 2012). A topic has a hierarchical structure of functional elements, i.e., elements which play certain communicative roles within the documents, and, at the highest level, is composed of the common elements: **Title**, **Short description**, **Prologue**, **Body** and **Related-links**.

According to topic types, DITA further defines more specific elements under the **Body** element. DITA provides by default several topic types such as *Concept topic*, *Reference topic* and *Task topic*. We focus here on the Task topic, which is designed for describing technical procedures, because what we are concerned is mainly municipal procedures. The left column in Table 1 shows the functional elements under the **Body** of Task topic (OASIS, 2010).

Note that the functional elements of the Task topic as defined in DITA are still too coarse-grained to properly organise municipal procedures and specify detailed linguistic patterns for each element. However, DITA allows for 'specialisation', so we undertook a genre analysis of actual municipal documents and assigned fine-grained sub-elements (the right column of Table 1).

2.2 Context-dependent CL with Pre-translation Processing

At this stage, the DITA structure provides a language-independent functional framework, which helps authors identify what information should be included. It is, however, still unclear how to write and translate each element. In order to instantiate the elements as texts, we defined context-dependent CL rules, i.e., desired linguistic patterns, for each element on both source and target sides.

For example, **Event condition** requires a conditional clause such as '日本に来たとき*/nihon ni kita toki*' (when you arrive in Japan). We also assigned a rather strict pattern for **Steps** element, polite speech style with declarative form 'します*/shimasu*' in Japanese and imperative form 'do' in English, such as '以下の書類を持参します*/ika no shorui o jizan shimasu*' (Bring the following documents), while the constraint is relaxed in **Context**, **Result** and **Postreq**.

The problem here is that a CL-compliant ST segment does not always generate a desired linguistic form in the target language. To resolve such incompatibilities, we introduce background pre-translation processing to transform the ST into an internal form amenable to the chosen MT system. Figure 1 depicts an example flow of this process for the **Steps** element: **ST1** is the CL-compliant original sentence, in polite speech style with the declarative '*shimasu*'. Since the MT output **MT1** is not a desirable result, **ST1** is transformed internally into **ST2**, with the imperative '*shiro*'. This then enables MT to produce

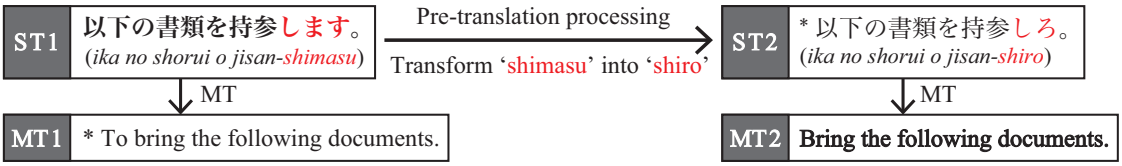

Figure 1: Pre-translation processing for **Steps** (* undesirable sentence)

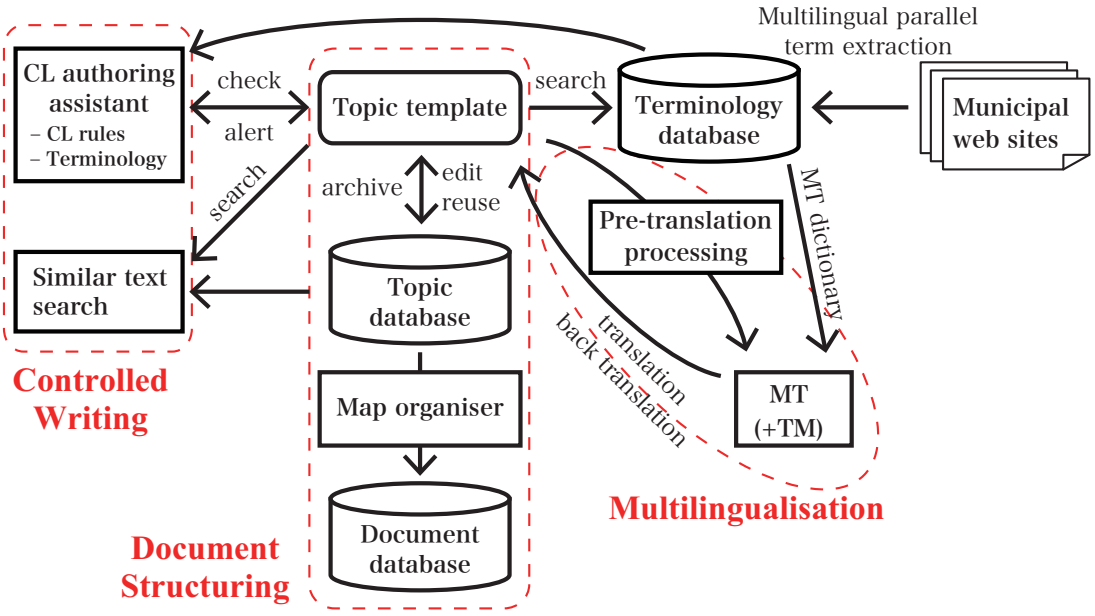

Figure 2: Modules of MuTUAL

MT2, which complies with the target side CL rule, with the use of the imperative form 'do'. **ST1** is served to Japanese readers and **MT2** to English readers. This process can be fully automated by defining simple transformation rules based on the morphological analysis of ST,[3] on condition that the linguistic patterns of the ST are sufficiently controlled in conjunction with functional elements.

3 The MuTUAL System

The MuTUAL system comprises modules for document structuring, controlled writing, and multilingualisation (see Figure 2). The following modules realise the contextual translation we have outlined:

- **Topic template** is the core interface for authoring self-contained topics in a structured manner. The left pane in Figure 3 provides the basic DITA Task topic structure for composing municipal procedural documents.
- **CL authoring assistant** analyses each sentence in the text box and highlights any segment that violates a local CL rule or controlled terminology, together with diagnostic comments and suggestions for rewriting (shown at bottom centre in Figure 3) (Miyata et al., 2016). In addition, we have implemented a preliminary rewriting support function with several of the features advocated by Mitamura et al. (2003). For a particular CL-noncompliant segment, the function offers alternative expressions; clicking one of the suggestions automatically replaces the offending segment in the text box above.
- **Pre-translation processing** automatically modifies source segments in the background following transformation rules defined for each functional element, and then **MT** produces the translation and back-translation at the same time.

[3]We used a Japanese morphological analyser MeCab. http://taku910.github.io/mecab/

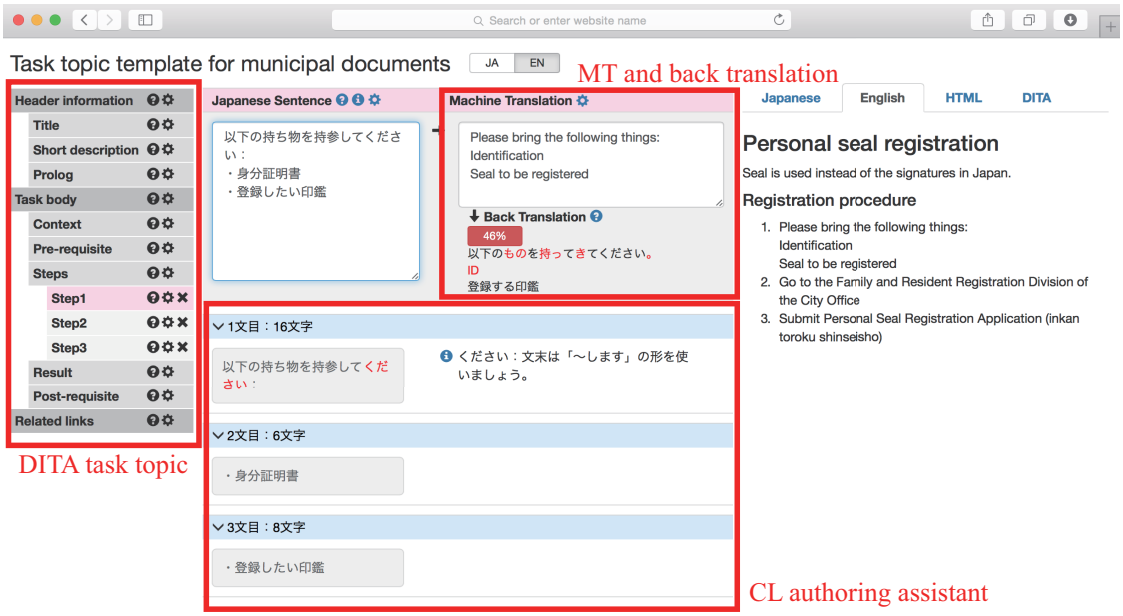

Figure 3: Task topic template for authoring municipal procedures

The key mechanism for enhancing authoring and translation is to invoke the CL authoring assistant tuned to the current functional element. For example, only for the **Steps** elements of the template, it implements the rule 'use declarative form *shimasu* at the end of the sentence'. Then the pre-translation processing for **Steps** transforms the declarative '*shimasu*' into an imperative '*shiro*' for our MT system to produce the desired imperative form 'do' on the target side.

The modules above are implemented in PHP and JavaScript, and can be accessed through the usual web browsers. The topic template seamlessly connects with our SMT system, TexTra,[4] the dictionary of which is customised by municipal terminology we maintain. We plan to publish an open-source version of the system through our project website.[5]

4 Conclusions and Future Work

We have presented a web-based support system for authoring municipal procedural documents. The principal novel feature of the system is that it makes use of document structuring based on the DITA framework, which affords a basis for fine-grained context-dependent CL rules coupled with pre-translation processing. It consequently enables MT to generate outputs appropriate to their functional context without degrading the quality of the source.

MuTUAL is currently operational online, focusing on the Task topic for creating municipal procedural documents in Japanese and English. The implemented CL rules were shown to be effective in triggering more appropriate outputs from our SMT system (Miyata et al., 2015). Also, a preliminary user evaluation revealed that the controlled authoring assistant module helped reduce time correcting CL-violations by more than 30%. As a future evaluation plan, while previous work has tended to focus on sentence-level text quality, we intend to evaluate the document-level quality of the system products by adopting task-based methods (Colineau et al., 2002). We will eventually make the system available to municipal departments and assess its usability in actual work scenarios.

[4]TexTra is a state-of-the-art SMT system particularly intended for Japanese as source language, and provides free API. https://mt-auto-minhon-mlt.ucri.jgn-x.jp

[5]The MuTUAL Project, http://www.mutual-project.com

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16J11185 and by the Research Grant Program of KDDI Foundation, Japan.

References

ASD. 2013. ASD Simplified Technical English. Specification ASD-STE100, Issue 6. http://www.asd-ste100.org.

Laura Bellamy, Michelle Carey, and Jenifer Schlotfeldt. 2012. *DITA Best Practices: A Roadmap for Writing, Editing, and Architecting in DITA*. IBM Press.

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation*, 16(3):175–218.

Nathalie Colineau, Cécile Paris, and Keith Vander Linden. 2002. An evaluation of procedural instructional text. In *Proceedings of the International Natural Language Generation Conference*, pages 128–135, New York.

Anthony Hartley and Cécile Paris. 2001. Translation, controlled languages, generation. In Erich Steiner and Colin Yallop, editors, *Exploring Translation and Multilingual Text production*, pages 307–325. Mouton, Berlin.

Richard Kittredge. 2003. Sublanguages and controlled languages. In Ruslan Mitkov, editor, *Oxford Handbook of Computational Linguistics*, pages 430–437. Oxford University Press, Oxford.

Tobias Kuhn. 2014. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170.

Teruko Mitamura, Kathryn Baker, Eric Nyberg, and David Svoboda. 2003. Diagnostics for interactive controlled language checking. In *Proceedings of the 4th Workshop on Controlled Language Applications (CLAW 2003)*, pages 237–244, Dublin, Ireland.

Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of the Machine Translation Summit XV*, pages 90–103, Miami, USA.

Rei Miyata, Anthony Hartley, Cécile Paris, and Kyo Kageura. 2016. Evaluating and implementing a controlled language checker. In *Proceedings of the 6th International Workshop on Controlled Language Applications (CLAW 2016)*, pages 30–35, Portorož, Slovenia.

OASIS. 2010. Darwin Information Typing Architecture (DITA) Version 1.2. http://docs.oasis-open.org/dita/v1.2/os/spec/DITA1.2-spec.html.

Sharon O'Brien and Johann Roturier. 2007. How portable are controlled language rules? In *Proceedings of the Machine Translation Summit XI*, pages 345–352, Copenhagen, DK.