氏　　名	梶原 智之（かじはら ともゆき）
所　　属	システムデザイン研究科　システムデザイン専攻
学 位 の 種 類	博士（工学）
学 位 記 番 号	シス博　第 101 号
学位授与の日付	平成 30 年 3 月 25 日
課程・論文の別	学位規則第 4 条第 1 項該当
学位論文題名	Text Simplification without Simplified Corpora （平易なコーパスを用いないテキスト平易化）
論文審査委員	主査　准教授　小町守 委員　教授　　山口亨 委員　准教授　高間康史 委員　准教授　岡崎直観　（東京工業大学大学院・情報理工学院）

【論文の内容の要旨】

Text simplification is the task of rewriting complex text into a simpler form while preserving its meaning. Systems that automatically pursue this task can potentially be used for assisting reading comprehension of less language-competent people, such as learners and children. Such systems would also improve the performance of other natural language processing applications, such as information extraction and machine translation. As with machine translation and abstractive summarization, this task is positioned as a Text-to-Text Generation task in natural language processing.

Current work has two approaches: lexical substitution and monolingual translation. In the former, a simpler synonymous sentence is generated by the pipeline of complex word identification, substitution generation, and substitution ranking. In the latter, a simpler synonymous sentence is generated using machine translation tools. In both approaches, mainstream methods acquire simplification rules from a large-scale parallel corpus. Therefore, text simplification was studied mainly in English for where rich resources are available. However, a large-scale simplified corpus for text simplification cannot be used in many

language other than English.

In this research, we propose text simplification methods by lexical substitution approach and monolingual translation approach for languages that cannot use large-scale simplified corpora, especially Japanese. As a lexical substitution approach without simplified corpora, we propose novel paraphrase acquisition, meaning preservation filtering, simplicity filtering, and grammaticality ranking methods for Japanese. Moreover, we build a first evaluation dataset for Japanese lexical simplification. In addition, as a monolingual translation approach without simplified corpora, we construct a pseudo-parallel corpus for text simplification from a raw corpus using readability assessment and sentence alignment, and enable text simplification using machine translation tools in any language.

Experimental results show that our lexical substitution approach outperforms the previous language-independent unsupervised method. Moreover, in the monolingual translation approach, the experimental results show that our sentence similarity measure achieved the state-of-the-art performance in alignment task of complex and simple sentences. In addition, our pseudo-parallel corpus succeeds in training machine translation tools as well as existing parallel corpora for text simplification.

Along with the above main works, we further improve text simplification through English experiments. First, we tackle the paraphrase acquisition task which is important for improving the lexical substitution approach. Next, we address the sentence similarity task which is important for improving the monolingual translation approach. Finally, we work on the quality estimation task for improving the automatic evaluation metrics of text simplification. With all these tasks, we achieved state-of-the-art performance.

The main contributions from this thesis are:

1. We propose state-of-the-art strategies for Japanese lexical simplification.
2. We build a first evaluation dataset for Japanese lexical simplification.
3. Our sentence similarity measure based on alignment between word embeddings outperforms previous works in alignment task of complex and simple sentences.

4. By improving sentence alignment, we achieve the best performance of English text simplification model using PBSMT.
5. For text simplification in languages that cannot use large-scale simplified corpora, we build a pseudo-parallel corpus from a raw corpus using readability assessment and sentence alignment. Experimental results show that our pseudo-parallel corpus can simplify as good as using large-scale simplified corpora.
6. We combine the paraphrasability score from monolingual corpora and from bilingual corpora to propose a novel paraphrasability score. we explained a well-known method for paraphrase acquisition, bilingual pivoting, as an unsmoothed version of PMI.
7. In order to further improve sentence alignment, we propose a domain adaptation method for sentence similarity measure based on alignment between word embeddings.
8. We improve a referenceless evaluation metric for simplification using word alignment.

The structure of this paper is as follows:

Chapter 2 provides a survey of the state-of-the-art in Text Simplification. We introduce both lexical substitution and monolingual translation approaches.

Chapter 3 presents our approach to Japanese lexical simplification. We build Japanese lexical simplification system (Contrib. 1) and evaluation dataset (Contrib. 2).

Chapter 4 presents our approach to English and Japanese sentence simplification. We investigate the best sentence alignment method for simplification (Contrib. 3) and build the state-of-the-art simplification model based on PBSMT (Contrib. 4). In addition, it shows that pseudo-parallel corpus obtained from a raw corpus by readability assessment and the sentence alignment is as effective as parallel corpus, and it opens the door to multilingualization of text simplification (Contrib. 5).

Chapter 5 presents experimental results in English for our three works to further improve text simplification. Section 5.1 considers both monolingual and bilingual corpus to acquire paraphrase lexicon more accurately (Contrib. 6). Section 5.2 proposes a domain adaptation method to calculate sentence similarity more accurately (Contrib. 7). Section 5.3 describes a quality estimation method using sentences similarity based on word alignment for more

accurate evaluation (Contrib. 8).

Finally, in Chapter 6, we provide our final remarks and directions for future work.

.