（ 続紙 1 ）

京都大学	博士（情報学）	氏名	Tolmachev Arseny
論文題目	Enhancing Morphological Analysis and Example Sentence Extraction for Japanese Language Learning （日本語学習のための形態素解析と例文抽出の高度化）		

（論文内容の要旨）

The goal of this thesis is to improve the state of foreign language learning by proposing a morphological analysis and example extraction approaches suitable for the task. During the process of learning foreign languages, students have to learn a large number of words, and example sentences greatly help the learning process. Automating the process of finding suitable example sentences is of great benefit to language learners. Still, because the Japanese language has a continuous script, the text must be segmented into words for further processing. In addition, educational applications require human-defined segmentation criteria. The process of segmenting text into words and inferring their parts-of-speech is called morphological analysis. The process of automated example sentence extraction is described in its chapter and two morphological analysis approaches are allotted a chapter each. Overall, the thesis consists of five chapters, including the introduction and conclusion.

The first chapter is the introduction. After giving the overall motivation for the work, it introduces the two types of approaches to morphological analysis: search-based and pointwise, giving a review for both of them. In the next section, the thesis gives motivation for example sentence extraction, defines the problem of high-quality example sentence extraction, and lists the related work.

The second chapter describes the improvements of the Juman++ morphological analyzer. It is a lattice-based morphological analyzer using a combination of trigram linear model and recurrent neural network models for scoring lattice paths. There were two types of improvements: algorithmic and microarchitectural. The most important algorithmic improvements are dictionary structure, hashing-based linear model feature computation framework, and trimming of beam search space. Microarchitectural improvements increased the efficiency of executing the Juman++ code. The improvements realized over 250 times analysis speed improvement over the original Juman++ version, while slightly improving the analysis accuracy. This chapter shows that a fully-lexicalized high-accuracy morphological analyzer utilizing a recurrent neural network-based language model can be practical.

The third chapter describes a neural network-based morphological analyzer in the pointwise paradigm, that uses only unigram characters as input. In contrast to lattice-based approaches, which utilize rich morphological dictionaries for the analysis, it is more difficult for pointwise approaches to utilize dictionary data in the analysis. The experiments show that when trained only on human-annotated data it cannot achieve high analysis accuracy. The thesis proposes to use a large

amount of automatically-analyzed data in addition to the human-annotated training examples. Using the automatically-annotated data by Juman++ achieves high analysis accuracy while having a very small model (15 MB compared to the 400 MB model of the lattice-based analyzer). The additional experiments show that it is possible to create even smaller models by sacrificing some accuracy.

The fourth chapter describes an approach for automated high-quality example sentence extraction for Japanese language learners. The approach, when given a target word, seeks to extract a set of example sentences each of which is good, and the set as a whole contains non-similar sentences. The proposed system consists of two components: a search engine and a selector. The search engine, in contrast to usual ones, can utilize syntactic and part of speech information and is used to find example sentence candidates which contain rich syntactic structure near the target word. The selector used a determinantal point process-based approach to select non-similar and high-quality sentences from the candidate list. The chapter also describes an evaluation experiment with Japanese language learners and a native teacher of Japanese. In the evaluation, the participants were asked to vote between three lists of example sentences, one for each method, on 14 lists for 14 selected target words. Both the learners and the teacher have preferred the proposed method to two baselines, especially for the intermediate (JLPT N3) level. The thesis also gives a detailed analysis of the reasons why the evaluators have selected examples extracted by a particular method.

The fifth chapter is the conclusion. It gives a review of the results of the thesis and discusses the remaining problems with the future work.

（続紙 ２ ）

（論文審査の結果の要旨）

外国語学習では多くの語を習得する必要があり、そこでは語の例文の提示が大きな助けとなる。これを計算機で支援する際、日本語のように単語が空白で区切られていない言語ではまず高精度な単語分割を実現する必要がある。本論文は、日本語学習の支援のための形態素解析の高度化と高品質な例文の抽出に関する研究成果をまとめたものである。得られた主要な成果は以下の通りである。

1. RNN(Recurrent Neural Network)言語モデルによって単語並びの意味的妥当性を考慮できる形態素解析器Juman++は、n-gram素性に基づく従来手法に比べて大幅に解析精度を改善したが、速度面では実用的ではなかった。本論文では、バイナリー辞書の構造の工夫、素性の計算の過程の改善、ビーム探索空間の削減によって、Juman++の解析速度を250倍以上に高速化し、実用化した。

2. 従来の高精度な形態素解析器は大規模なモデルサイズを要するものであった。本論文では、ニューラルネットワークによる文字単位の分類に基づく省スペースかつ高精度な形態素解析モデルを考案した。人手でタグ付けされた訓練データでモデルを学習するだけでは低精度に留まるが、大規模な自動解析結果を合わせてモデルを学習することにより従来モデルと同等の高精度な解析を実現した。この時、モデルサイズは従来モデルに比べて20分の1以下であった。また、解析精度を犠牲にすることで、モデルサイズをさらに削減することが可能であることを示した。

3. 外国語学習支援で重要となる高品質な例文の自動抽出システムを提案した。当該システムは構文構造と品詞情報を用いる検索システムと、多様性を重視する決定的点過程に基づく選択アルゴリズムからなり、与えられた語に対して高品質かつ似通っていない例文集合の自動抽出を実現した。23名の日本語の学習者と日本語母語話者の教師による評価実験の結果、２つのベースライン手法と比較して優位性を示した。特に、日本語能力試験のN3（中級）レベルの学習者から高い評価が得られた。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和４年２月１６日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。また、本論文のインターネットでの全文公表についても支障がないことを確認した。

要旨公開可能日：　　年　　月　　日以降