（ 続紙 1 ）

京都大学	博士（情報学）	氏名	Graham NEUBIG
論文題目	Unsupervised Learning of Lexical Information for Language Processing Systems （言語処理システムのための語彙的情報の教師なし学習）		

（論文内容の要旨）

Natural language processing systems such as speech recognition and machine translation conventionally treat words as their fundamental unit of processing. However, in many cases the definition of a “word” is not obvious, such as in languages without explicit white-space delimiters, in agglutinative languages, or in streams of continuous speech.

This thesis attempts to answer the question of which lexical units should be used for these applications by acquiring them through unsupervised learning. This has the potential to lead to improvements in accuracy, as it can choose lexical units flexibly, using longer units when justified by the data, or falling back to shorter units when faced with data sparsity. In addition, this approach allows us to re-examine our assumptions of what units we should be using to recognize speech or translate text, which will provide insights to the designers of supervised systems. Furthermore, as the methods require no annotated data, they have the potential to remove the annotation bottleneck, allowing for the processing of under-resourced languages for which no human annotations or analysis tools are available.

Chapter 1 provides an overview of the general topics of word segmentation and morphological analysis, as well as previous research on learning lexical units from raw text. It goes on to discuss the problems with the existing approaches, and lays out the general motivation for and techniques used in the work presented in the following chapters.

Chapter 2 describes the overall learning framework adopted in this thesis, which consists of models created using non-parametric Bayesian statistics, and inference procedures for the models using Gibbs sampling. Non-parametric Bayesian statistics are useful because they allow for automatically discovering the appropriate balance between model complexity and expressive power. We adopt Gibbs sampling as an inference procedure because it is a principled, yet flexible learning method that can be used with a wide variety of models. Within this framework, this thesis presents models for lexical learning for speech recognition and machine translation.

With regards to speech recognition, Chapter 3 presents a method that can learn a language model and lexicon directly from continuous speech with no text. This is performed using the hierarchical Pitman-Yor language model, a non-parametric Bayesian formulation of standard language modeling techniques based on the Pitman-Yor process, which allows for principled and effective modeling and inference. With regards to modeling, the non-parametric formulation allows for learning of appropriately-sized lexical units that are long enough to be useful, but not so long as to cause sparsity problems. Inference is performed using Gibbs sampling with dynamic programming over weighted finite states transducers (WFSTs). This

makes it straight-forward to learn over lattices, allowing for language model learning in the face of acoustic uncertainty. Experiments demonstrate that the proposed method is able to reduce the phoneme error rate on a speech recognition task, and is also able to learn a number of intuitively reasonable lexical units.

In the work on machine translation, Chapter 4 presents a model that, given a parallel corpus of sentences in two languages, aligns words or multi-word phrases in each sentence for use in machine translation. The model is hierarchical, allowing for the inclusion of overlapping phrases of multiple granularities, which is essential for achieving high accuracy when using the phrases in translation. Inference is performed using Gibbs sampling over trees expressed using inversion transduction grammars (ITGs), a particular form of synchronous context-free grammar that allows for the expression of reordering between languages and polynomial-time alignment through the process of biparsing. Experiments show that this model is able to achieve translation accuracy that is competitive with the process used in traditional systems while reducing the model to a fraction of its original size.

Chapter 5 extends this model to perform alignment over multi-character substrings, learning a model that directly translates character strings from one language to another. In order to do so, two changes are made to improve alignment. The first improvement is based on aggregating substring co-occurrence statistics over the entire corpus and using these to seed the probabilities of the ITG model. The second improvement is based on introducing a look-ahead score similar to that of A* search to the ITG biparsing algorithm, which allows for more effective pruning of the search space. An experimental evaluation finds that character-based translation with automatically learned units is able to provide comparable results to word-based translation while handling linguistic phenomena such as productive morphology, proper names, and unsegmented text.

Chapter 6 concludes the thesis with an overview of the task of lexical learning for practical applications and directions for future research.

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、音声認識や機械翻訳といった自然言語処理を行うシステムにおいて、従来先見的に与えられていた「単語」などの単位を、教師なしのノンパラメトリックベイズ学習の枠組みで自動的に獲得する研究をまとめたものであり、得られた主な成果は次の通りである。

1. 連続音声と音素認識器から語彙と言語モデルを自動獲得する枠組みを提案し、その方法を定式化した。これは、音素認識と単語区切りの曖昧性を各々表現したラティスを合成してできる WFST (Weighted Finite State Transducer) に対して、ギブスサンプリングを行うことで実現される。本手法により、妥当な語彙を獲得して言語モデルを構築し、音素認識率を改善できることを示した。

2. 二言語の対訳コーパスから、機械翻訳のためのフレーズの単位を階層的に自動獲得する方法を定式化した。これは、ITG (Inversion Transduction Grammar) による解析木を階層的に表現したものに対してギブスサンプリングを行うことで実現される。本手法により、従来法と比べて、翻訳の際に用いるフレーズの数を 20%以下に削減しながら、同等の翻訳精度を実現することができた。

3. 上記では単語列からフレーズを抽出していたが、これを文字列に適用できるように拡張と効率化を行い、単語辞書を与えなくても、従来の単語を単位とする機械翻訳システムと同程度の翻訳精度を実現できることを示した。

以上のように本論文は、日本語を含む、単語間の区切りのない言語や膠着言語においても、従来先見的に定義されていた単語などの単位に関して、タスク・データに応じた最適化を行う方法を提示するもので、学術上・実際上寄与するところが少なくない。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、平成 24 年 2 月 22 日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。