論 文 内 容 の 要 旨

博士論文題目 Learning Lexical Representations for Neural Machine Translation
（邦題： ニューラル機械翻訳のための単語表現学習）

氏 名 Philip Arthur

要 旨

Words are the most prominent unit in languages. Human started to understand languages by combining some characters forming the smallest unit of languages that convey meaning. Since babies, human naturally learn words by combining a sequence of speech. It is very important to model this correctly on computers. Until recently, there have been several different representations of words in a computer. For example, the earliest of natural language processing treated words as atomic units. While simple and effective, this approach has several drawbacks such as it creates a data sparsity problem because words that should be correlated to each other are now treated as totally different tokens. As a remedy to this approach is the continuous representation which treated words as a vector in a continuous space, thus better representing the sense of meanings of the word itself for a computer.

As it is aforementioned, this thesis studies the better lexical representation for a computer, particularly in doing translations. Wrong representation of words unit can make learning slower, or even worse, fail to generalize the pattern of the languages. This generalization failure can cause NMT system to fail to acquire/produces certain lexical units. We first study how to help Neural Machine Translation system to produce better rare lexical units by directly adding some probability priors to the system. Next, we study how to model these most basic units in a continuous space so the systems can do better in modeling the language. Finally, we apply an unsupervised method to extract words unit from a stream of unsegmented input, mimicking how babies do language acquisition. There are mixed results for the whole experiments. We successfully increase the system performance in generating lexical units using a lexicon. We also achieved a better state of the art performance by using the mix of character and word representation in continuous space. However, in the third experiment, we achieved partially good results. While we are successful in acquiring a good knowledge of a language such as segmentation, our experiments still show that it is increasing the translation performance. Further studies are needed to stabilize the proposed method.

氏　名	Philip Arthur

（論文審査結果の要旨）

This thesis studies better word representation of words for neural machine translation (NMT). Most natural language processing (NLP) work so far treated word as the most basic unit that convey meaning. This approach is simple and effective but also has several drawbacks due to disparity of counts and availability of words in the corpus, which is used by NLP algorithms for language acquisition. This effect is also seen in a language in which one word can occur more often than the others and causes *rare words problems*. This thesis pursues better lexical representation of words, specifically focusing on the continuous representation in the task of end-to-end NMT. The first study proposed the better way of representing the output representation in the continuous space. This work uses the goodness of the count-based previous MT systems to provide a strong prior for the NMT. The second study investigated the better way of representing the input representation of NMT. This work tries to (1) compare each of the lexical unit combination, (2) combine the composition functions together. Experiments show that the composition function that uses the character bag-of-ngrams information gained the best accuracy. The third study tries to remove the limit of the fixed size vocabulary units on the input side by trying to discover the lexical units during the training of NMT. This work models the segmentation as hidden variables that are discovered jointly in the training of NMT. It employs an unsupervised method of reinforcement learning using translation quality as the reward.

The research proposed solutions to the problems which haven't been solved and series of his research resulted in one journal paper and two peer reviewed international conference papers. As a result, the thesis is sufficiently qualified as Doctoral thesis of Engineering.