論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Improving Word Representations for Language Modeling
著者(和文)	FENGYukun
Author(English)	Yukun Feng
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12394号, 授与年月日:2023年3月26日, 学位の種別:課程博士, 審査員:奥村 学,熊澤 逸夫,中山 実,篠﨑 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12394号, Conferred date:2023/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	Engineering Information and Communications Engineering	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	(Engineering)
学生氏名： Student's Name	Feng Yukun		指導教員（主）： Academic Supervisor(main)	Manabu Okumura	
			指導教員（副）： Academic Supervisor(sub)		

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

Language modeling (LM) is an important task in the natural language processing field, with various applications such as speech recognition, machine translation and summarization. Recently, neural language models (NLMs) have shown a great success and are better than traditional count-based methods. Standard NLMs usually maintain a fixed vocabulary and map each word to a continuous representation. These models cannot handle out-of-vocabulary words and are also not effective for learning the relationships between words for infrequent words. For example, although the words ``husbandman'' and ``salesman'' share the suffix ``man'' in their surface forms, standard NLMs cannot capture such information in obtaining the relationship between the two words. One solution is to use smaller units, such as bytes, characters, or word pieces learned from word tokens. However, this approach has to process longer sequences than word-based alternatives and may increase modeling and computational challenges. In addition, some previous work indicated that in some cases subwords are notably worse than word-level models with subword-awareness in neural machine translation tasks. Another solution to deal with these issues is to use character-level information of each word to calculate the word representation, and it is often referred to as character-aware NLMs. However, there are still two research questions for these models. First, neither character-aware NLMs nor standard NLMs are effective for learning the semantic relationship of infrequent words, such as “innumerable” and “myriad” as they do not share any surface form. Second, although character-aware NLMs make use of character-level information, it is still common to inject the word-level information together, as it provides information from a different aspect. Thus, how to effectively inject word-level information in character-aware NLMs becomes a research topic.

In this thesis, we propose a simple and effective usage of word clusters applied to Continuous Bag-of-Words (CBOW), which can produce enhanced word embeddings for improving NLMs regarding the first research question. These word clusters consist of words that function similarly and are useful for data sparsity. Particularly, many word clustering algorithms can be applied to a raw corpus with different languages to help us obtain word clusters easily without additional language resources. In our method, we keep only very frequent words and replace the other words with their clusters for both input and output words in the CBOW model. This is motivated by the fact that word clusters are more reliable than infrequent words. Thus, only very frequent word embeddings and a small amount of cluster embeddings are produced as the output. We apply these learned word embeddings and cluster embeddings to the fine-tuning of NLM tasks. At the beginning, the embeddings of infrequent words within one cluster are initialized by the same embedding of their cluster and are then updated differently in accordance with their context. In our experiments, we choose two standard NLM benchmarks and two machine translation (MT) tasks for evaluating our proposal. To investigate the effect of word clusters across different languages, eight typologically diverse languages are further selected for the LM task. Finally, we also analyze our proposal in detail, such as the effect of different word clustering algorithms, the gain of our proposal for infrequent and frequent words, speed and spatial comparison, as well as the effectiveness of our proposal on large-scale corpus.

Regarding the second research question, previous work usually inject word-level information at the input side of NLMs through a gating mechanism, or averaging or concatenation of word vectors. Because these approaches generally target at the input vectors, the word-level information is not explicitly taken into account at the output layer for predicting the next word and thus these methods may not make full use of word-level information. To deal with this problem, we propose to inject the information of current and previous words at the output layer. Our method can be viewed as a combination of a modern character-aware

NLM and a simple n-gram word-level language model. This is strongly inspired by the success of n-gram language models. In our experiments, we selected 14 datasets with typologically diverse languages. We show that our injection method is better than previous methods that inject word-level information at the input, and our method can be also used together with these previous injection methods. Finally, we also focus on analyzing the effectiveness of the word-level information in character-aware NLMs and our injection method applied to them in various languages. For the effectiveness of word-level information, we analyzed the effects of rare words and what kind of words work best when injected. For analyzing our injection method, we tested several variants of our injection method, such as injecting character-level information or combination of word- and character-level information into the output layer. These comparisons can reveal more properties of our injection method used in character-aware NLMs.

（博士課程）
Doctoral Program

東京工業大学
Tokyo Institute of Technology