様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　６ 月 １１ 日現在

機関番号：	１２６０８
研究種目：	若手研究(B)
研究期間：	2017～2018
課題番号：	１７Ｋ１２７３９
研究課題名（和文）	Corpora on Demand: Scalable Methods of Obtaining Linguistic Data
研究課題名（英文）	Corpora on Demand: Scalable Methods of Obtaining Linguistic Data

研究代表者

ドローズド　アレクサンドロ（Drozd, Aleksandr）

東京工業大学・情報理工学院・研究員

研究者番号：９０７４０１２６

交付決定額（研究期間全体）：（直接経費）　　1,400,000円

研究成果の概要（和文）：テキスト表現のための評価ベンチマークのセットを開発しました。テキストを収集するためのスクリプトを開発しました。これにより、クーポラの品質を評価できます。すべての結果(ソフトウェアとリソース)は　http://vecto.space/ プロジェクトウェブサイトで公開されています。ソフトウェアはすべてオープンソースです。

研究成果の学術的意義や社会的意義
This research project contributed to the development of methods of filtering textual data and evaluating text representations.

研究成果の概要（英文）：We have developed a set of evaluation benchmarks for text representations. These benchmarks can be used to estimate quality of text corpora, as well as methods for constructing representations themselves and corresponding hyperparameters. We have implemented heuristics for filtering good quality text fragments and instrumental scripts to help parsing different formats of archived internet documents. Since internet documents might contain copyrighted and private information, we could not publish the raw corpora. Instead we are publishing all our source codes to extract raw texts from different archive formats as well as models (such as word embeddings) that we have trained on large scale texts. All resources and source codes are available at the project website http://vecto.space

研究分野： natural language processing

キーワード： corpora　evaluation　text representations

様　式　C−１９、F−１９−１、Z−１９、CK−１９（共通）

１．研究開始当初の背景

At the beginning of research project most of experiments in computational linguistics and natural language processing were done on a couple of easy-to-get corpora, mainly dumps of Wikipedia articles. Another problem is that NLP models were often compared using different corpora and on a small set of probing tasks.

２．研究の目的

The purpose of this research project was to help NLP and Cl practitioners to get training data easily and evaluate how particular choice of training corpora affects performance of downstream tasks.

３．研究の方法

In this research project we had to address both scientific and engineering challenges. We integrated existing and created new methods for filtering text and evaluation of text representations. Besides observing metrics like accuracy for evaluation tasks, we were also trying to investigate which linguistic phenomena are involved in the pipeline of each task and how downstream results reflect what upstream stages like model and source corpus capture in terms of linguistic nuances. We were trying to find and integrate existing solutions when possible and have made ourselves software implementations for novel ideas, including those proposed by us in a context of this project.

４．研究成果

We have developed novel methods for evaluating text representations, including new algorithms to solve analogical reasoning task, new datasets for Japanese, English and Russian languages and reference models for various diagnostic tasks. These evaluations methods can be used to benchmark various stages of NLP pipelines, including choice or methods to create source corpora. We have created an approach to preserve metadata at all stages of NLP pipelines, so that impact of individual stages can be easily evaluated. We have created helper scripts to extract raw texts from sources like Wikipedia, project Gutenberg collection of books and Common Crawl archives. We have integrated some of existing solutions for detecting language, parsing HTML pages etc.
We also investigated linguistic-theoretical issues related to collection of textual data and evaluation of text representation and pre-trained model in transfer-learning scenarios.
We have integrated all our results in an open-source software library called Vecto. The library and resources are available at the http://vecto.space/ website. In Vecto library we dedicated special attention to the problem of reproducibility of results. We have implemented metadata collection and assimilation in all stages of supported NLP pipelines, including choice and collection of source corpora. Resources we published include pre-trained embeddings, reference implementation of evaluation models, datasets etc. User-base of vecto is continuously growing: we are regularly receiving feature requests, bug reports etc.
We could not publish the corpora we collected to avoid risk of copyright violation complains. Instead we have released models and embeddings which were pre-trained using corpora we have collected.
We have organized a tutorial (“Compositionality in the Age of Word Embeddings”, collocated with LREC 2018 conference) and promoted some of our results and software in it. We have also organized a mini-symposium on convergence of HPC and AI, co-located with SIAM PP 2018 conference, in which one of the sessions was dedicated to the problem of scaling NLP models.
Collaborations established during this research continue after the expiration of the funding and we aspire to extend and improve existing results.

５．主な発表論文等

〔雑誌論文〕（計　　件）

〔学会発表〕（計　4　件）

1) Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. Subcharacter information in japanese embeddings: when is it worth it? In In Proceedings of the Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP (RELNLP) 2018, 28–37. ACL, 2018.

2) Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. Subword-level composition functions for learning word embeddings. In Proceedings of The 2nd Workshop on Subword and Character level models in NLP (SCLeM), 38–48. ACL, 2018.

3) Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. Investigating different syntactic context types and context representations for learning word embeddings. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2421–2431. 2017.

4) Anna Rogers, Aleksandr Drozd, and Bofang Li. The (too many) problems of analogical reasoning with word vectors. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017), 135–148. 2017.

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

http://vecto.space

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：