https://lib.uliege.be https://matheo.uliege.be

Master's Thesis : NetBERT: A Pre-trained Language Representation Model for Computer Networking

Auteur : Louis, Antoine
Promoteur(s) : Louppe, Gilles
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2019-2020
URI/URL : http://hdl.handle.net/2268.2/9060

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

University of Liège - Faculty of Applied Sciences
In collaboration with Cisco Systems

NetBERT: A Pre-trained Language Representation Model for Computer Networking

Author:
Antoine Louis

Advisor:
Prof. G. Louppe

Obtaining accurate information about products in a fast and efficient way is becoming increasingly important at Cisco as the related documentation rapidly grows. Thanks to recent progress in natural language processing (NLP), extracting valuable information from *general domain* documents has gained in popularity, and deep learning has boosted the development of effective text mining systems. However, directly applying the advancements in NLP to *domain-specific* documentation might yield unsatisfactory results due to a word distribution shift from general domain language to domain-specific language. Hence, this thesis aims to determine if a large language model pre-trained on domain-specific (*computer networking*) text corpora improves performance over the same model pre-trained exclusively on general domain text, when evaluated on in-domain text mining tasks.

To this end, we introduce NetBERT (Bidirectional Encoder Representations from Transformers for Computer Networking), a domain-specific language representation model based on BERT and pre-trained on large-scale computer networking corpora. Through several extrinsic and intrinsic evaluations, we compare the performance of our novel model against the general-domain BERT. We demonstrate clear improvements over BERT on the following two representative text mining tasks: networking *text classification* (0.9% F1 improvement) and networking *information retrieval* (12.3% improvement on a custom retrieval score). Additional experiments on *word similarity* and *word analogy* tend to show that NetBERT capture more meaningful semantic properties and relations between networking concepts than BERT does. We conclude that pre-training BERT on *computer networking* corpora helps it understand more accurately domain-related text.

Master of Science in Data Science & Engineering

Academic year 2019-2020