Semantic Entity Retrieval Toolkit

Van Gysel, C.; de Rijke, M.; Kanoulas, E.

Publication date
2017

Document Version
Author accepted manuscript

Published in
Neu-IR: Workshop on Neural Information Retrieval

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date: 10 Mar 2024
Semantic Entity Retrieval Toolkit

Christophe Van Gysel
University of Amsterdam
Amsterdam, The Netherlands
cvangysel@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

Evangelos Kanoulas
University of Amsterdam
Amsterdam, The Netherlands
e.kanoulas@uva.nl

ABSTRACT
Unsupervised learning of low-dimensional, semantic representations of words and entities has recently gained attention. In this paper we describe the Semantic Entity Retrieval Toolkit (SERT) that provides implementations of our previously published entity representation models. The toolkit provides a unified interface to different representation learning algorithms, fine-grained parsing configuration and can be used transparently with GPUs. In addition, users can easily modify existing models or implement their own models in the framework. After model training, SERT can be used to rank entities according to a textual query and extract the learned entity/word representation for use in downstream algorithms, such as clustering or recommendation.

KEYWORDS
Neural information retrieval; Entity retrieval; Toolkit

1 INTRODUCTION
The unsupervised learning of low-dimensional, semantic representations of words and entities has recently gained attention for the entity-oriented tasks of expert finding [9] and product search [8]. Representations are learned from a document collection and domain-specific associations between documents and entities. Expert finding is the task of finding the right person with the appropriate skills or knowledge [1] and an association indicates document authorship (e.g., academic papers) or involvement in a project (e.g., annual progress reports). In the case of product search, an associated document is a product description or review [8].

In this paper we describe the Semantic Entity Retrieval Toolkit (SERT) that provides implementations of our previously published entity representation models [8, 9]. Beyond a unified interface that combines different models, the toolkit allows for fine-grained parsing configuration and GPU-based training through integration with Theano [3, 6]. Users can easily extend existing models or implement their own models within the unified framework. After model training, SERT can compute matching scores between an entity and a piece of text (e.g., a query). This matching score can then be used for ranking entities, or as a feature in a downstream machine learning system, such as the learning to rank component.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGIR 2017 Workshop on Neural Information Retrieval (NeuIR'17), August 7–11, 2017, Shinjuku, Tokyo, Japan
© 2017 Copyright held by the owner/author(s).
class ExampleModel(VectorSpaceLanguageModelBase):
    def __init__(self, *args, **kwargs):
        super(ExampleModel, self).__init__(*args, **kwargs)

        # Define model architecture.
        input_layer = InputLayer(
            shape=(self.batch_size, self.window_size))
        ...

        def loss_fn(pred, actual, _):
            # Compute symbolic loss between
            # predicted/actual entities.
            # The framework deals with underlying boilerplate.
            self._finalize(loss_fn, _____)

        def get_representations(self):
            # Returns the representations and parameters
            # to be extracted.

    # Get representations of the examples.

Code snippet 1: Illustrative example of the SERT model interface. The full interface supports more functionality omitted here for brevity. Users can define a symbolic graph of computation using the Theano library [6] in combination with Lasagne [3].

2.2 Representation learning

After the collection has been processed and packaged in a machine-friendly format, representations of words and entities can be learned. The toolkit includes implementations of state-of-the-art representation learning models that were applied to expert finding [9] and product search [8]. Users of the toolkit can use these implementations to learn representations out-of-the-box or adapt the algorithms to their needs. In addition, users can implement their own models by extending an interface provided by the framework. Code snippet 1 shows an example of a model implemented in the SERT toolkit where users can define a symbolic cost function that will be optimized using Theano [6]. Due to the component-wise organization of the toolkit (Fig. 1), modeling and text processing are separated from each other. Consequently, researchers can focus on modeling and representation learning only. In addition, any improvements to the collection processing (§2.1) collectively benefits all models implemented in SERT.

2.3 Entity ranking & other uses of the representations

Once a model has been trained, SERT can be used to rank entities w.r.t. a textual query. The concrete implementation used to rank entities depends on the model that was trained. In the most generic case, a matching score is computed for every entity and entities are ranked in decreasing order of his score. However, in the special case when the model is interpreted as a metric vector space [2, 8], SERT casts entity ranking as a k-nearest neighbor problem and uses specialized data structures for retrieval [5]. After ranking, SERT outputs the entity rankings as a TREC-compatible file that can be used as input to the trec_eval1 evaluation utility.

Apart from entity ranking, the learned representations and model-specific parameters can be extracted conveniently from the models through the interface2 and used for downstream tasks such as clustering, recommendation and determining entity importance as shown in [10].

3 CONCLUSIONS

In this paper we described the Semantic Entity Retrieval Toolkit, a toolkit that learns latent representations of words and entities. The toolkit contains implementations of state-of-the-art entity representations algorithms [8, 9] and consists of three components: text processing, representation learning and inference. Users of the toolkit can easily make changes to existing model implementations or contribute their own models by extending an interface provided by the SERT framework.

Future work includes integration with Pyndri [11] such that document collections indexed with Indri can transparently be used to train entity representations. In addition, integration with machine learning frameworks besides Theano, such as TensorFlow and PyTorch, will make it easier to integrate existing models into SERT.

Acknowledgments. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This research was supported by Ahold Delhaize, Amsterdam Data Science, the Bloomberg Research Grant program, the Criteo Faculty Research Award program, the Dutch national program COMMIT, Elsevier, the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Google Faculty Research Award scheme, the Microsoft Research Ph. D. program, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nr 612.001.116, H0R-11-10, CI-14-25, 652.002.001, 612.001.551, 652.001.003, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES