JF17K Data Formats

Jianfeng Wen, Yongyi Mao and Richong Zhang

Document Created @ 2016-04-26 11:36 EST

JF17K are extracted and filtered from Freebase for the study of knowledge base embedding involving non-binary relations. The process of creating the dataset is described in our paper *On the representation and embedding of knowledge bases beyond binary relations* (IJCAI 2016). This document contains the description of the data formats of JF17K, where the terminologies in that paper are used in this document.

Overall, the data are presented in three different formats, stored separately in the following three folders

- `instances/`
- `instancesWithFactID/`
- `S2C_triples/`

Folder `instances/`

This folder contains training set $\mathcal{G}^{\checkmark}$ (in `train.txt`) and testing set $\mathcal{G}^{?}$ (in `test.txt`). Each file in the folder is explained below.

- `entity.txt` contains the list of all entities included in the dataset; each line specifies one entity.
- `relation.txt` contains the list of all relation types included in the dataset together with their fold (or arity) values; each line contains two space-delimited fields, `relationType` and `foldValue`.
- `train.txt` contains the list of all training instances. Each line specifies one instance. For each J-fold relation, its roles are taken as integers $\{1, 2, \ldots, J\}$. If the instance is from a J-fold relation, the line specifying the instance contains $J + 1$ tab-delimited fields; the first field is `relationType`, and the remaining J fields are in order `entity4Role_1`, `entity4Role_2`, . . . , `entity4Role_J`.
- `test.txt` contains the list of all test instances, where each line specifies one instance. The format of `test.txt` is nearly identical to that of `train.txt` except that in each line, there is an additional field at the beginning specifying the index (ID) of the test instance.

Folder `instancesWithFactID/`

This folder contains training set $\mathcal{G}_{\mathrm{id}}^{\checkmark}$ (in `train.txt`) and testing set $\mathcal{G}_{\mathrm{id}}^{?}$ (in `test.txt`). The data in this folder is identical to that in folder `instances/` except that for instances contained in the same fact, the fact ID of the instances are kept. The files in the folder are explained below.

- `entity.txt` contains the list of all real entities (the entities in folder `instances/`) and the fact-ID entities; each line specifies one entity.
- `real_entity.txt` contains the list of all real entities; each line specifies one entity.

- relation.txt contains the list of all relation types together with their fold values. Each line contains two space-delimited fields, relationType and foldValue. Note that for some relation types, the fold value is one greater than that in the same relation type in folder instances/. This is because FACT-ID is included as an additional role of the relation.
- train.txt contains the list of all training instances, where each line specifies an instance. The format of the file is the same as that of train.txt in folder instances/.
- test.txt contains the list of all test instances, where each line specifies an instance. The format of the file is the same as that of test.txt in folder instances/.

Folder S2C_triples/

This folder contains training set $\mathcal{G}^{\checkmark}_{\text{s2c}}$ (in train.txt) and testing set $\mathcal{G}^{?}_{\text{s2c}}$ (in test.txt). The datasets are obtained by applying the S2C conversion to $\mathcal{G}^{\checkmark}$ and $\mathcal{G}^{?}$ respectively.

The files in the folder are explained below.

- entity.txt contains the list of all entities included in the dataset; each line specifies one entity.
- relation.txt contains the list of all relation types included in the dataset together with their fold values; each line contains two space-delimited fields, relationType and foldValue. Note that the relation types here are not the original relation types in Freebase (or in $\mathcal{G}^{\checkmark}$ and $\mathcal{G}^{?}$). They result from the S2C conversion. As such, all relations here are binary, and the fold values are all equal to 2.
- train.txt contains the list of all training triples. Each line specifies one triple and each triple is written in three tab-delimited fields in order of relationType, headEntity, and tailEntity. We note that for each relation type, the two roles, HEAD and TAIL, are chosen arbitrarily and this choice is used consistently for all triples having this relation type.
- test.txt contains the list of all tes triples, where each line specifies one triple. The format of test.txt is nearly identical to that of train.txt except that in each line, there is an additional field at the beginning specifying the index (ID) of the test triple.