K-Graph Completion

Author
osanchez@umass.edu

1 Problem statement

Our goal for this project is to show the effectiveness of using TransH over TransE for knowledge graph completion. Ultimately we want to see if transH is better than TransE and what drawbacks, if any, arise when using it. Our original plans were to create a knowledge graph and use both translations to see if there were any differences or improvements in query results, but after research conducted on how to implement a knowledge graph from scratch, we found that resources were limited. Any knowledge graph information we found online were in reference to Google's knowledge graph API or were completed implementations done by individuals with minimal documentation. We had started an implementation by first gathering pre-processed data from the SQuAD or BaBi dataset, but later found that the FB15K and WN18 data sets were closer to what we needed. We also realized that for the time frame on this project, it was not possible for our current level of knowledge on the topic to create a knowledge graph from scratch. As an alternative, we decided that we would use a knowledge graph with implemented translations from the following github repository: KB2E. This repository includes a knowledge graph implementation using the data we initially planned to use. It also contains an implementation for each of our translating models that we want to compare.

2 Your dataset

Our project has multiple files that are used to construct the knowledge graph and train and test each translation model. The files are pre-processed from the FB15k and the WN18 data sets. The files are as follows:

train.txt - Training file, format (entity 1, entity 2, rel)
valid.txt - Validation file, same format as train.txt
test.txt - Test file, same format as train.txt
entity2id.txt - All entities and corresponding ids, one per line
relation2id.txt - All relations and corresponding ids, one per line

Looking at each file we could see that we were working with a fair amount of data. the train file, which is used to train each of the models, contains nearly 16,000 entity pairs and relationships. a sample of this data is show below (Figure 1). The validation file contains two entities as well as their relationship. It is a file that contains actual entity relationships that the model will be tested against to see if the results returned for each translation method are giving acceptable precision and recall values.The test file contains 2 entities and a relationship which will be replaced by entities in the knowledge graph, and later be ranked in descending order of similarity scores. entity2id and relation2id are files that map entity ids found in the freebase data dump files to integer values. One of the biggest issues we came across was that the freebase knowledge base that the data had been constructed from had been discontinued. we were able to create a script that uses Google's new Knowledge Graph Search API to translate the entity ID's found in each file to real entities to better analyze the data and results.

3 Baselines

The transE knowledge graph model was created. The model is trained using the data found in the train.txt file. In TransE, relationships are represented as translations in an embedding space: if a relation between entities hold, then the embedding of the object entity should be close to the em-

bedding of the subject plus some vector representing the relationship. Farther away relations signify that the relationship is not strong.

Below are the parameters used to test the model and the epoch phase and loss value:

size = 100
learning rate = 0.001
margin = 1
method = bern
relation_num=1345
entity_num=14951

epoch:0 169462
epoch:1 64286.4
...
epoch:999 608.396

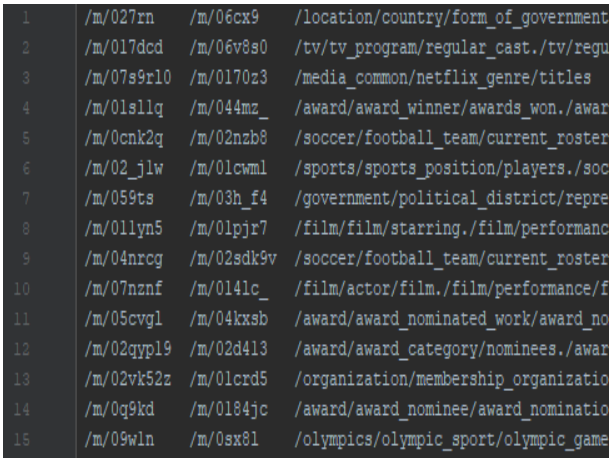
4 Error analysis

The data we are working with contains a file called valid.txt. This contains entity pairs as well as their relationship. Once our models are trained we will test them using the values in the test.txt file. valid.txt contains the true entity relationship for each pairing so we will add a hit or miss function that indicates whether or not the model predicted close to accurate results or not. We also want to create 3D graphs of entity relationships at various different epochs for each model and with the script we created to get information on the entities, we will also see if any of the relationship matching make sense to us.

5 Your approach

We currently have a trained TransE model. A majority of our issues are running tests on the trained model. We are currently analyzing the code (written in c++ and python) trying to figure out what exactly our results mean. One major issue we had was that the freebase knowledge base had been discontinued. Because of this we were unable to get information on each entity ID, so the data didn't have much meaning to us. We were able to create a script using the Google Graph Search API to retrieve entity information on these ID's. While some of the ID's do seem to have been removed from the knowledge base (or given a new ID), we believe that we can continue using our data. The model was created using a desktop equipped with an Intel i7-8000k CPU and a overclocked GTX 1080 ti GPU so we have more than enough processing power.

6 Figures



1	/m/027rn	/m/06cx9	/location/country/form_of_government
2	/m/017dcd	/m/06v8s0	/tv/tv_program/regular_cast./tv/regu
3	/m/07s9rl0	/m/0170z3	/media_common/netflix_genre/titles
4	/m/01sl1q	/m/044mz_	/award/award_winner/awards_won./awar
5	/m/0cnk2q	/m/02nzb8	/soccer/football_team/current_roster
6	/m/02_j1w	/m/01cwm1	/sports/sports_position/players./soc
7	/m/059ts	/m/03h_f4	/government/political_district/repre
8	/m/01lyn5	/m/01pjr7	/film/film/starring./film/performanc
9	/m/04nrcg	/m/02sdk9v	/soccer/football_team/current_roster
10	/m/07nznf	/m/0141c_	/film/actor/film./film/performance/f
11	/m/05cvgl	/m/04kxsb	/award/award_nominated_work/award_no
12	/m/02qyp19	/m/02d413	/award/award_category/nominees./awar
13	/m/02vk52z	/m/01crd5	/organization/membership_organizatio
14	/m/0g9kd	/m/0184jc	/award/award_nominee/award_nominatio
15	/m/09w1n	/m/0sx81	/olympics/olympic_sport/olympic_game

Figure 1: A sample of data taken from the train.txt file. The data shows two entities in the first and second column and their relationship listed in the third.

7 Timeline for the rest of the project

- Complete TransE and TransH models by 11/24
- Analyze the output of each model, do an error analysis by 11/28
- Complete final report and submit poster by 12/6
- Prepare presentation by 12/10 for presentation on 12/11

8 References

Lin, Yankai, et al. "Learning entity and relation embeddings for knowledge graph completion." AAAI. Vol. 15. 2015.

Vandenbussche, Pierre-Yves. Pierre-Yves Vandenbussche. PierreYves Vandenbussche, 29 Aug. 2017, pyvandenbussche.info/2017/translating-embeddings-transe/.