情報工学実験 II

1 先週したこと

1.1 TransE の動作確認

いただいた TransE のサンプルコードの動作確認をした. どのような処理をしているかの雰囲気を理解した.

1.2 Datasets, Dataloaders の理解

Dataset, Dataloader とは, PyTorch においてサンプルデータを扱う基本要素のこと. Dataset にはサンプルとそれに対応するラベルが格納される. また, Dataloader にはイテレート処理が可能なデータが格納される. Dataloader はサンプルを簡単に利用できるように Dataset をイテレート処理可能なものへとラップする. イテレート処理とは, 反復や繰返しをする処理のこと.

1.3 MRR の理解

情報検索システムの評価指標のひとつ. MRR (Mean Reciprocal Rank) では, 検索結果のランキングにノイズが少なく, できるだけ上位に適合文書が存在することを重視するタイプの指標で, 特に, 検索結果ランキングを上位から眺めたときに何番目に適合した文書 (正解) があるか確認する. ある検索結果において, 最初に見つかった適合文書のランクの逆数をスコアとする（ ex. 3 位 = 1 / 3 ）. 得られた検索結果に対するスコアを複数そろえたときに平均としたものが MRR 値となる.

$$0.0 \leq MRR \leq 1.0$$

$MRR > 0.5$ のとき, 平均的に 2 件目までに適合文書があることが期待できるため, 一定の性能があることがわかる.

また, MRR が扱う適合度は, 適合か不適合かの 2 値（ binary relevance ）が前提となっており, 部分適合等の段階的な適合度（ graded relevance ）を扱うことはできない.

2 今日の目標

2.1 ナレッジグラフ, TransE の理解

ナレッジグラフとは, さまざまな知識を体系的に連結し, その関係をグラフ構造で表した知識のネットワークのことである. 主に Web を知識源として知識を収集し, 集めたデータを 3 つ組構造へと変換することによりさまざまなデータのつながりを柔軟に表現することが可能である. この 3 つ組構造は「もの - 関係 - もの」や「 S - V - O 」などのように表される. ナレッジグラフ推論チャレンジでは, 「場面（シーン）」とそれに対応する「主語（ subject ）」「述語（ predicate ）」「目的語（ what, whom, when, where, etc. ）」「場面間の関係（ if, then, because, etc. ）」「絶対時間」「場面の原文」で表現されている.

TransE とは, グラフを 1 つのベクトル空間に落とし込む Gragh embedding を行う手法のことである. 入力としてナレッジグラフを与えるとすべての head, relation, tail に対してそれぞれ k 次元のベクトルを出力する. つまり, TransE を用いることでナレッジグラフをベクトル空間上に embedding することができる. これにより, ベクトルを用いた単語同士の加算をすることが可能になる.

$$(head) + (relation) = (tail)$$

2.2 [head, relation, tail] の tensor 作成

[head, relation, tail] の 3 つ組を entity と relation の 2 つに分けてそれぞれの id を作成した. そして, その 3 つ組を id の 3 つ組へと変換し, リストとして保存した. そのリストを tensor に変換した.

3 来週までにすること

3.1 Dataset の作成

[head, relation, tail] の 3 つ組を id 化して作成した tensor を用いて Pytorch の Dataset を作成する. また, Dataloader の動作も確認する.

4　来週の目標

4.1　TransE モデルの学習

作成した Dataset を用いて TransE モデルに学習させる. これによって, head, および relation, tail のうち 1 つを予測するモデルを作成する.