情報工学実験 II Report

1　授業日の内容

2 値分類のためのデータ前処理

用意されているデータに対して 2 値分類を行うために前処理を行った．「0[tab]’ 文章’」のようにしてデータセットを作成した．改行で split 関数を用いると文章がない場合もデータセットに含まれるため, その処理に苦労した．

テーマ決め

ナレッジグラフについて実験することを決め, 今後何をするかを話し合った. 扱うデータセットをもらい, それをどのように処理していくかを決めた．

2　1 週間の内容

2 値分類のためのデータ前処理の続き

苦労した改行の処理をインターネットを駆使して解決した．その後の BERT への処理は引き続き考える．

データ前処理と id への変換

受け取った HDF5 ファイルにあるデータセットにおいて, タイトルを除去するという前処理を行った．その後, そのデータセットの entity の id を作成した．この使い方をまだ知らないため, 今後この使い方を学んでいく．

必要な知識

TransE や ConceptNet について調べた．

TransE とは, Gragh embedding を行う手法のことで, 入力として Knowledge gragh を与えるとすべての subject, predicate, object に対してそれぞれ k 次元のベクトルを出力する．

ConceptNet とは, Open Mind Common Sense (OMCS) を用いて作成された常識的知識ベースの生成プロジェクトである．また, 単語がノードによって関係付けられており, ある単語の背景にある知識や常識を提供する多言語ナレッジグラフである．

3　今後行うこと

前処理を行ったデータを用いた実行

受け取ったデータセットに前処理を行ったデータセットを用いて, その後の処理を施していく．まず, ConceptNet を用いたデータセットの拡張を行う. そして, 次に何をすればよいかの説明を受け, プログラムを作成する．

また, 2 値分類に関しても, どのようにすれば BERT を用いて動作するかを考え, 実行する．

ナレッジグラフについて理解を深める

ナレッジグラフについてまだ理解できていないところが多々あるため, ナレッジグラフだけでなくその周辺の知識も調べて学習する．