Combining Contextual Words and Knowledge Graph Embeddings

Software project, work update 2

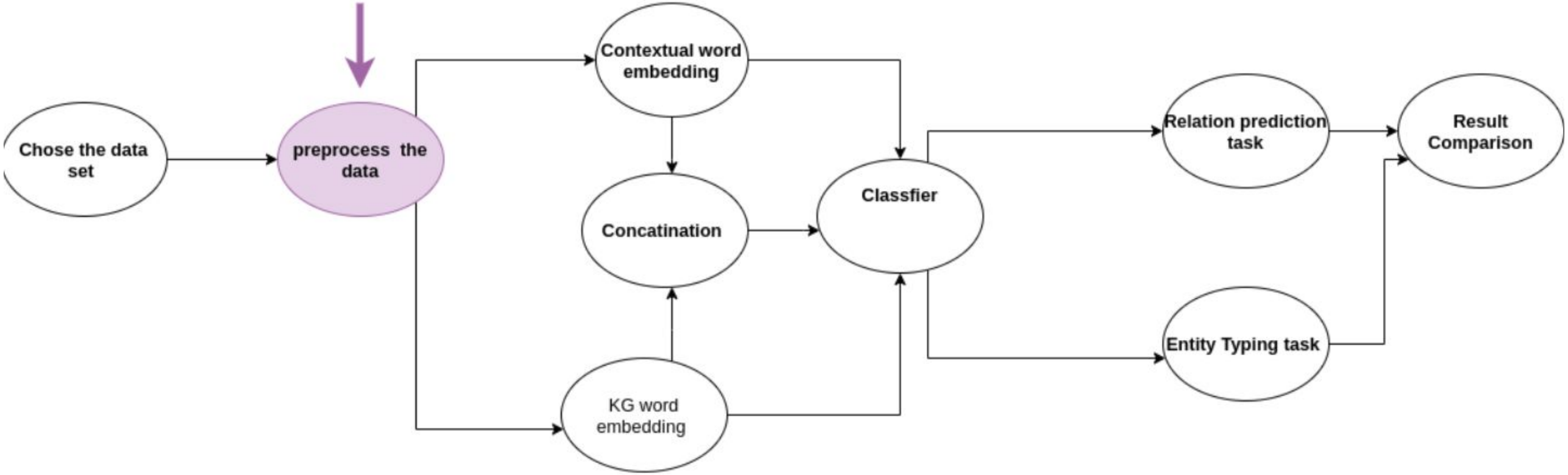
IDMC, University of Lorraine

6/10/2020

OUTLINES

1. Contextual Word Embedding
2. Current state
3. Knowledge Graph Embedding
4. Concatenation model and combining embedding method
5. New dataset: PGx Corpus
6. MilEstone and Future direction

Current State :

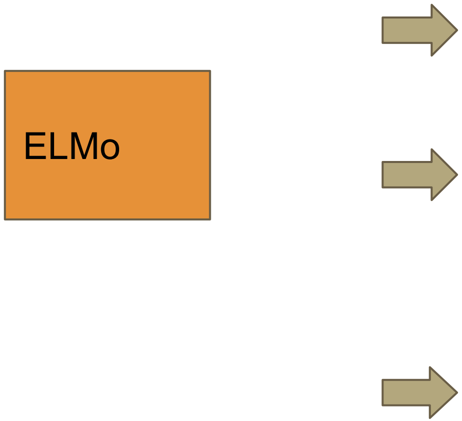


CONTEXTUAL WORD EMBEDDING

Contextual Word Embedding :

- Using pre-trained **ELMo**(Embeddings from Language Models) model.

- Its features :



ELMo

Contextual: The representation for each word depends on the entire context in which it is used.

Deep : The word representations combine all layers of a deep pre-trained neural network.

Character based: allowing the network to use morphological clues to form robust representations for the out-of-vocabulary tokens unseen in training.

ELMo Model:

- This model is pre-trained with a self-supervising task called a bidirectional language model.

- ELMo pretrained models are trained on **Google 1-Billion Words dataset**, which was tokenized with the **Moses Tokenizer**.

- ELMo gained its language understanding from being trained to predict the next word in a sequence of words - a task called *Language Modeling*

CONCATENATION MODEL AND COMBINING EMBEDDING METHOD

Combination of Contextual word embedding and KG embeddings :

- Last Year experiment [1] :
  - “The model simply concatenates the embedding generated with the two pre-trained models we use for KB and contextual data (BigGraph and ELMo respectively).”
  - No removing or adding .
- Our way for combining the two embeddings approaches : in the future work section .

KNOWLEDGE GRAPH EMBEDDING

ENTITY AND RELATION EMBEDDINGS FOR KNOWLEDGE GRAPH

Knowledge Graphs encode structured informations and their rich relations

- Predict relations between entities under supervision of the existing KG
- Nodes in KB are different types and attributes
- Edges in KB are relations of different types

BIG GRAPH : TOOL TO CREATE LARGE GRAPH EMBEDDINGS

Embedding system to that incorporates several modifications to traditional multi-relation embedding :

- A block decomposition of the adjacency matrix into N buckets
- A distributed execution model
- Efficient negative sampling for nodes that samples negative nodes both uniformly and from the data
- Support for multi-entity multi-relation graphs with perrelation configuration options such as edge weight and choice of relation operator

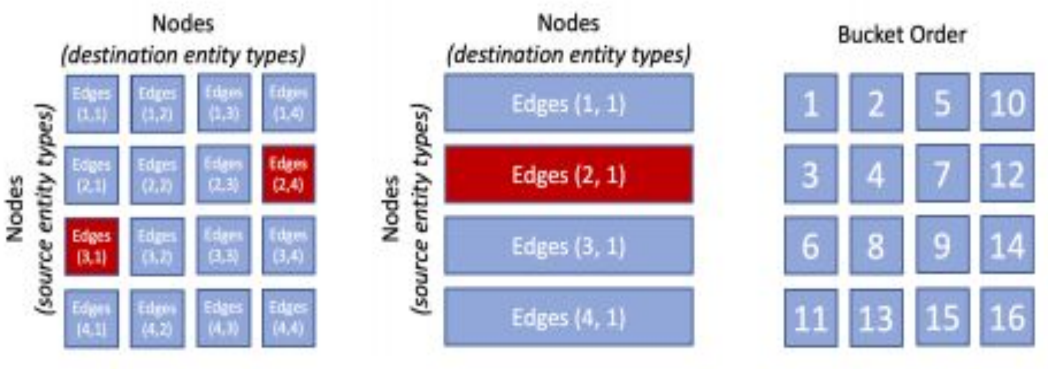
BIG GRAPH MODEL

A multi-relation graph is a directed graph G = (V, R, E) :

V = nodes

R = set of relations

E = a set of edges



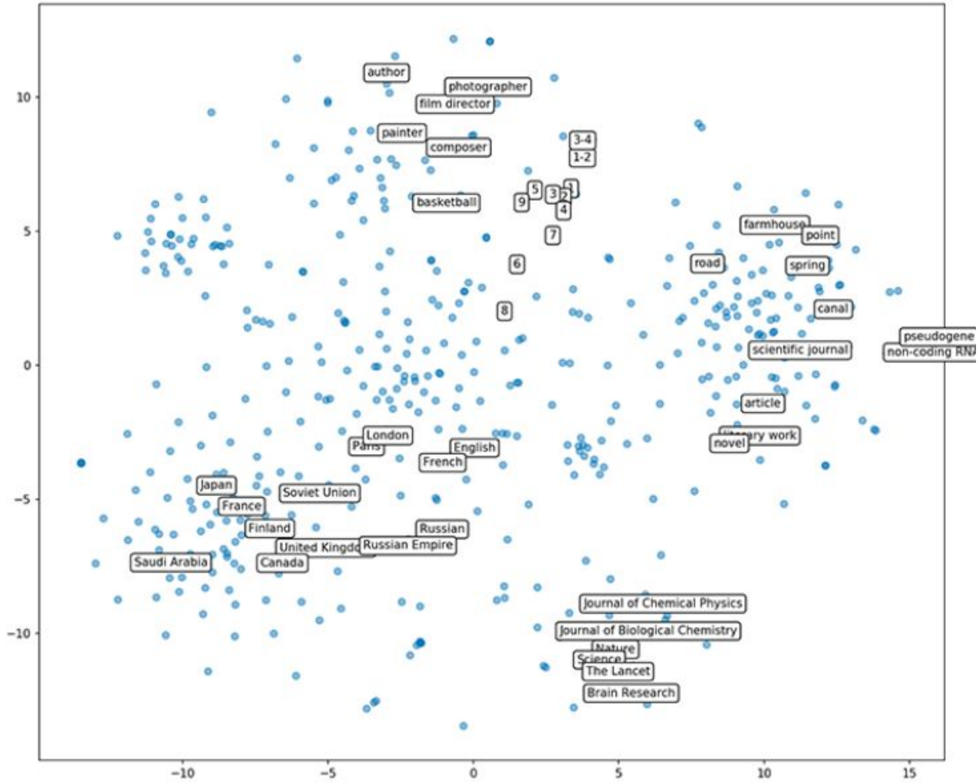
The PBG partitioning scheme for large graphs.

Left : nodes are divided into partitions. Edges are divided into buckets based on the partition of their source and destination.

Central : Entity types with small cardinality do not have to be partitioned

Right : the ‘inside-out’ bucket order guarantees that buckets have at least one previously-trained embedding partition

Evaluating pytorch-BigGraph



- FB15k Dataset
- 5,000 nodes and 600,000 edges

A data set of this size can fit on a modern server, but PBG's partitioned and distributed execution reduces both memory usage and training time

A t-SNE plot of some of the embeddings trained by PBG for the Freebase knowledge graph. Entities such as countries, numbers, and scientific journals have similar embeddings.

NEW DATASET: PGx CORPUS

Pharmacogenomics (PGx) Dataset

- Comprises 945 sentences from 911 PubMed abstracts.
- Manually annotated relations.
- Annotated with PGx entities of interest (mainly gene variations, genes, drugs and phenotypes),
- and relationships between those.

Datasets

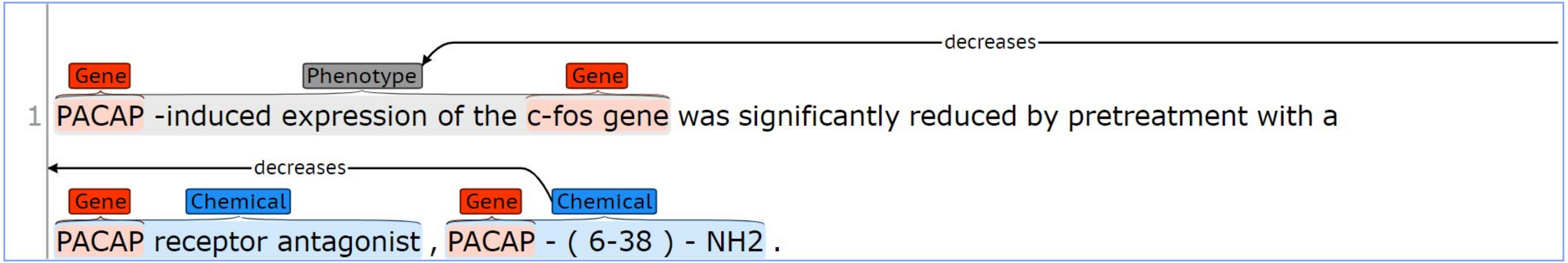
Entity Types	798
Relationships	16
Sentences	13,874
No. of Relations	29,492

Freebase-NYT

Entity Types	10
Relationships	7
Sentences	945
No. of Relations	2,871

PGx Corpus

Example



PACAP -induced expression of the c-fos gene was significantly reduced by pretreatment with a PACAP receptor antagonist , PACAP - (6-38) - NH2 .

Sample

Proenkephalin gene expression in the primate uterus : regulation by estradiol in the endometrium .

Annotation

```
T1  Phenotype 0 29 Proenkephalin gene expression
T2  Gene_or_protein 0 13   Proenkephalin
T3  Chemical 68 77 estradiol
R1  influences Arg1:T3 Arg2:T1
```

Advantages of PGx Corpus

- Manual annotation of relationships.
- Less Preprocessing.
- Coarser classification of entities.

Drawbacks of PGx Corpus

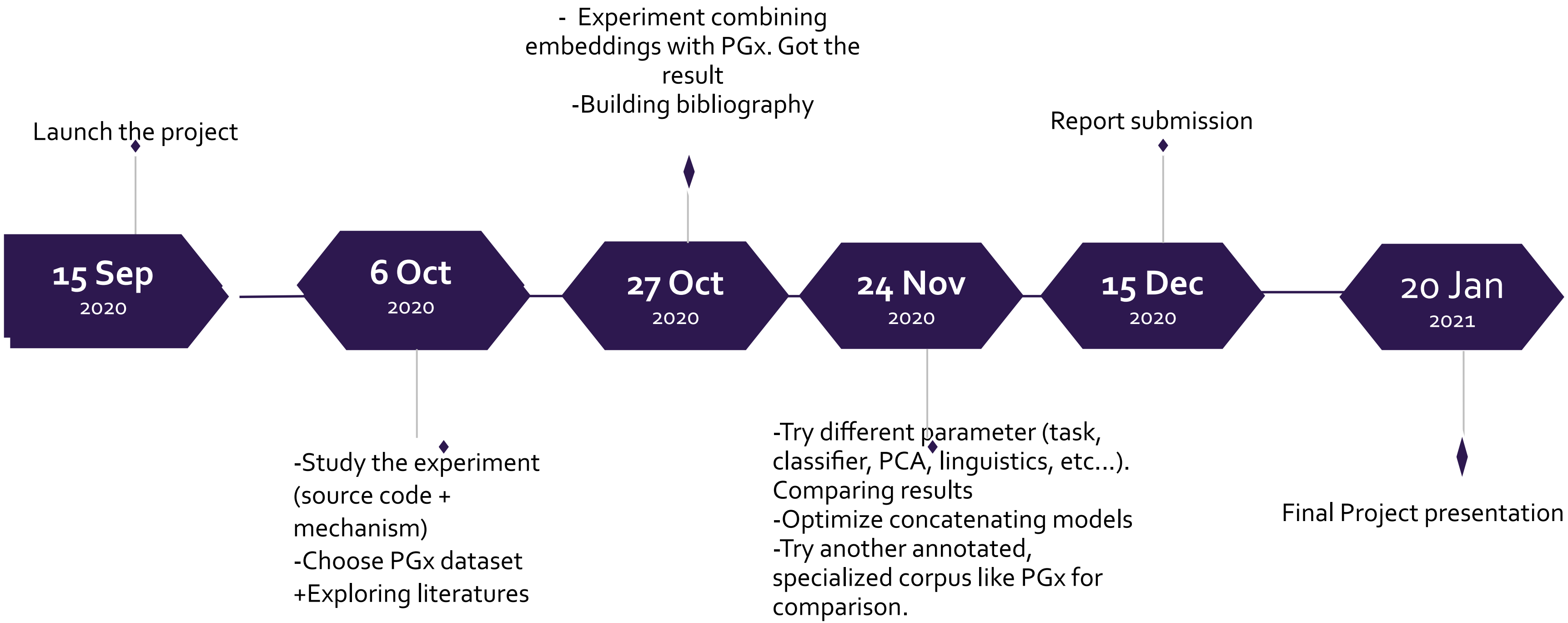
- Less Data.

First-hand information of PGx Corpus

- Transfer Learning
- Integrate with larger, common knowledge Corpus
- Analysis of the role that syntactic features may play in TL.
- Problem of syntactical formation
- Apply Neural networks without a large corpus

FUTURE DIRECTION AND MILESTONES

WORKFLOW



15 Sep
2020

Launch the project

6 Oct
2020

-Study the experiment (source code + mechanism)
-Choose PGx dataset
+Exploring literatures

27 Oct
2020

- Experiment combining embeddings with PGx. Got the result
-Building bibliography

24 Nov
2020

-Try different parameter (task, classifier, PCA, linguistics, etc...). Comparing results
-Optimize concatenating models
-Try another annotated, specialized corpus like PGx for comparison.

15 Dec
2020

Report submission

20 Jan
2021

Final Project presentation

PROPOSED DIRECTIONS

- Proposed Corpus:

-TACRED: same function as Freebase but larger than its subsets (15K, NYT)

-SemEval: Large training corpus for semantic analysis, with specialized subset like DrugBank

- Changing parameters of the experiments: PCA for embedding, syntactic analysis, experiment with various classifiers.
- Optimising the Concatenation Model.

References :

1. Dieudonat, Léa & Han, Kelvin & Leavitt, Phyllicia & Marquer, Esteban. (2020). Exploring the Combination of Contextual Word Embeddings and Knowledge Graph Embeddings.
2. https://allennlp.org/elmo
3. https://towardsdatascience.com/introduction-to-pytorch-biggraph-with-examples-b50ddad922b8#:~:text=PyTorch%20BigGraph%20is%20a%20tool,it%20to%20a%20neural%20network.&text=And%20then%20use%20it%20as%20features%20in%20a%20traditional%20neural%20network
4. https://arxiv.org/pdf/1903.12287.pdf
5.

Thank you!