**Project name:** Find Hidden Entities in Wikipedia Articles

**Dataset:** Wikidata5m

**Reference:** Bidirectional relation-guided attention network with semantics and knowledge for relational triple extraction.

**Purpose:**

* Extract relationship triples from Wikipedia articles
* Build a neural network and train it to extract the triples (head, relationship, tail) from unstructured text
* Knowledge graph that we extract where head representing usually the subject and tail representing the object

**Approach for preparation:**

We are using wikidata5m dataset which contains three files containing the data:

1. Transductive: triples of the knowledge graph where exists triples in the shape (head id, relation id, tail id)
2. Corpus: containing Wikipedia articles and each one is represented by id
3. Entity & relation aliases: containing for each query\_id a list of aliases representing it and for each relationship\_id a list of names for the relationship

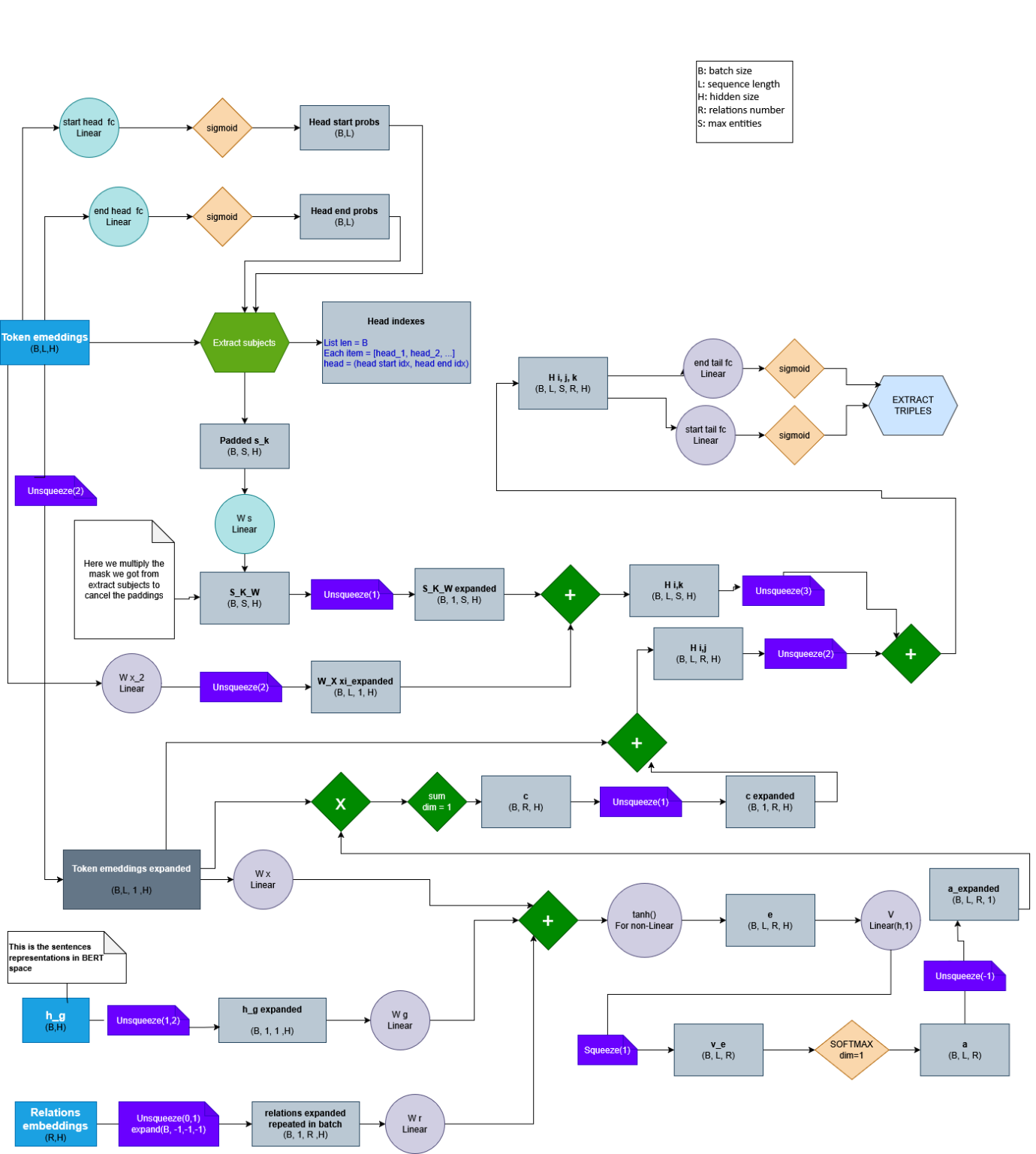
**BRASK approach:**

BRASK aims to extract correct triples, doing 2 extractions (forward & backward) and take the intersection of those triples, this approach will solve also overlapping problems

The bidirectional extraction let us use attention guidance in both directions:

1. Semantic general language knowledge for the forward extraction representing general knowledge and based on BERT
2. Domain knowledge (knowledge graph) for the backward extraction where we train transE on our existing knowledge graph

**Neural network graph:**



**Preparation:**

Before we start the training we did some preparations which I will talk about, every step of preparation saves the result of preparation in a .pkl file and all the pkl files paths are saved in Data.py file.

Preparation:

1. Prep.py

Wikidata5m dataset has three main files:

1. wikidata5m\_text.txt containing the Wikipedia text for each entity
2. wikidata5m\_alias.tar containing aliases for the texts, and for the relationships
3. wikidata5m\_transductive.tar containing knowledge graph (triples) of the dataset

In prep.py we are converting them into dictionaries for easy access in next steps and saving the results in dictionaries.

The result dictionaries are:

1. Description texts: including the entity\_id and the entity\_description (entity text)
2. Aliases: including the entity\_id and the entity\_aliases (list of strings or titles referring to the entity\_id)
3. Relationships: including relationship\_id and relationship\_names (list of strings)
4. Triples – knowledge graph: including triples (head\_id, relationship\_id, tail\_id) where the head\_id & tail\_id referring to the entity\_id in 1&2 dictionaries and relationship\_id referring to 4

Also, we are creating [10,100,1k, 1m, full ] dictionaries from those 4 dictionaries to be able to test the model with min results. When we create min and let’s say 10, we will have:

* 10 descriptions chosen randomly
* Get the triples of those 10 descriptions
* Get triples of those 10 descriptions
* Add descriptions of the tails to the descriptions dictionary
* Get relationships of those triples
* Get aliases of all descriptions

The descriptions will not be 10 but will be 10 + descriptions of the triples’ tails

A black and white screen with white lines

AI-generated content may be incorrect.

1. Normalize.py

In this file we normalize the descriptions dictionary, we do this by parallel, so if we work with the full dataset (5m descriptions), it would be more performant to normalize

The normalization of the description will do the following into the texts:

1. Replace special chars with their English version using re compiled patterns
   1. For example ắặằẳẵǎâấậầ would be converted to ‘a’
   2. The preparation of the compiled pattern to do this step is done in file prep.py
2. Remove any non-english chars (Chinese, Arabic… )
3. Remove multiple white spaces

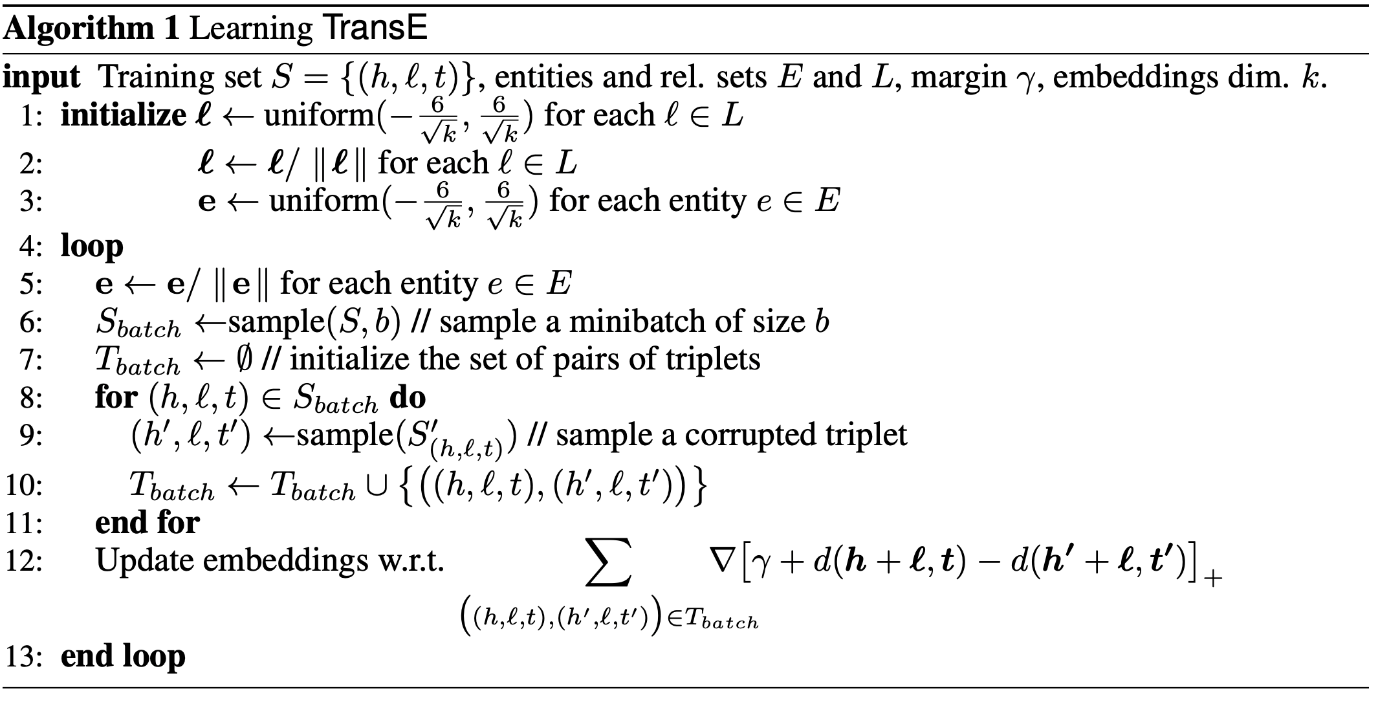
We did not do any more normalization (like removing punctuation marks, make the description small letters) to not loose the meaning of the sentences.

A grey rectangular object with black border

AI-generated content may be incorrect.

1. TransE.py

**TransE** is an energy-based model that produces knowledge base embeddings. It models relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities.



The basic idea is that if we have a triple (head, relation, tail), and each one of head,relation and tail are represented in some space, if you take the vector of head + vector of relation, you should arrive near the vector of the tail (+- some margin)

We use transE to translate our knowledge graph into an embeddings, we will take the relationship embeddings and represent them as our knowledge attention guidance when we do backward extraction.

The code in transE.py uses the triples dictionary and train transE neural network to generate rel\_embeddings which we will use in the backward extraction.

I did 140 epochs with 52224 as the batch\_size and run the model on cuda and it results in perfect training for the neural network and saved the resulted relationship embeddings.

1. Prep\_rels.py:

In this file, we are preparing our relationship embeddings for forward extraction based on BERT last 2 hidden states.

This is representing the semantic knowledge of our model.

The steps we are doing to get embeddings of our relations:

1. Extract relations used in triples dictionary
2. Loop through each relation
3. Get aliases of this relation (strings)
4. Get the last two hidden layers of each string representing the relation (more general)
5. Take the average pool of those embeddings to have an embedding for the relation

Also, there is an optimized version of the function extracting the relation embeddings for the gpu use

1. Prep\_ground\_truth.py

In this file, we are preparing our silver spans to calculate loss while training.

Basically what we have in our dataset is triples for each head.

For example let’s say we have 2 descriptions (entities) (texts) in our descriptions dictionary:

1. “q1” 🡪 Europe is a continent of the world continents
2. “q2” 🡪 Italy is a country in Europe

And we have the following aliases:

1. “q1” 🡪 [“europe”, “oropa”, “eu”]
2. “q2” 🡪 [“Italy”, “italia” , “it”]

And we have those triples:

1. “q1” 🡪 [ (), … other triples ]
2. “q2” 🡪 [(q2, r1, q1), …other triples] (considering r1 is relationship for “located\_in”)

I have triple (q2, r1, q1) representing that “Italy is located in Europe”

From this triple, we have to create silver spans for the q2 which will show where the head (Italy) starts and ends, and where is the tail (Europe) starts and ends.

Notes:

* for simplicity this example head’s and tail’s are one token but they might be 5 tokens
* for simplicity we are showing a raw text, in real life, we are tokenizing the sentences so one token here might be multiple tokens after tokenization using BERT tokenizer)

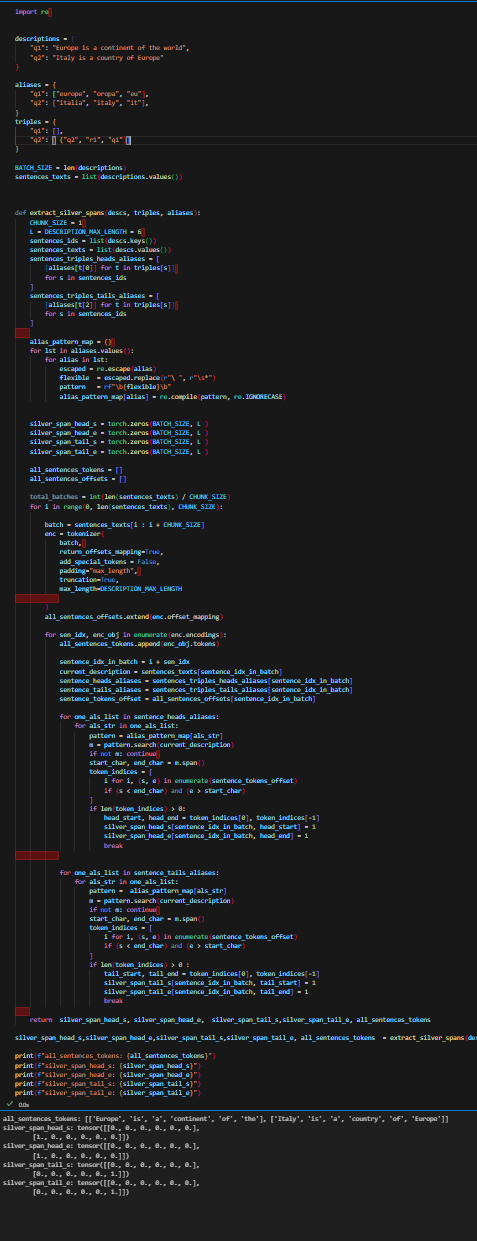
We try to search for the head/tail aliases in the description and we decide the start/end index for each one of them.

Figure - Code screenshot for running the example on test notebook and executing the function extract\_silver\_spans -

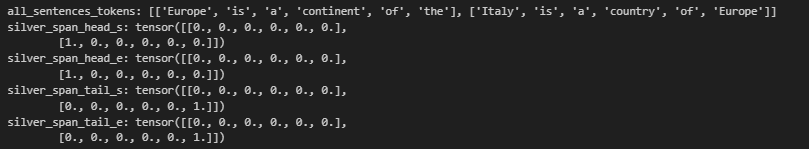
In our example, the triple is (q1, \_, q2), the aliases for head are [“europe”, “oropa”, “eu”] and the aliases for tail are [“Italy”, “italia” , “it”], we can find from our description text “Italy is a country in Europe”, that the head “italy” starting at index 0 and ending at index 0, and for our tail “europe” starting at index 5 and ending at index 5 (considering here tokenization is based on space split).

We know that for q2, the triple (q2, \_, q1):

* Head’s start index = 0
* Head’s end index = 0
* Tail’s start index = 5
* Tail’s end index = 5

We are using re patterns to search for the alias string in the description.

Because in our model, we are doing binary tagging system in our subject/object extractions, we have to create silver spans tensors that each one have the shape (batch\_size, sequence\_length), where each value in those tensors is 1 if the token is (head\_start, head\_end, tail\_start, tail\_end)



The code gave result about the tokens (fortunately for this small example BERT tokenizer is not tokenizing multiple tokens for one word).

But we notice:

* silver\_span\_head\_s[1][0] is 1 because Italy is the start of head
* silver\_span\_head\_e[1][0] is 1 because Italy is the end of head
* silver\_span\_tail\_s[1][-1] is 1 because Europe is the start of tail
* silver\_span\_tail\_e[1][-1] is 1 because Europe is the end of tail

1. The model:

Dataset:

The dataset that we will prepare for our model is a dataset including the description dictionary (only ones that have triples, for full 5m descriptions, we found triples in 4.8m), we will batch those descriptions along with their silver spans (head start, head end, tail start, tail end) into batches through dataloader. Also, we will extract h\_g for each description which is representing the embedding of the text, we will get this by pool average of the tokens inside the text using BERT.



Inside the forward of the model we will do:

1. Create probabilities of the start/end of the subjects (head) for the forward extraction, start/end of object (tail) for the backward extraction:

A close-up of a math problem

AI-generated content may be incorrect.

This will be generated by feeding the token embeddings into fully connected neural network to get probabilities of distinguishing each token in the input sequence.

Then from those probabilities we will get the spans (0,1) for those probabilities:

A computer screen shot of white text

AI-generated content may be incorrect.

Each probability tensor has shape (Batch\_size, sequence length)

Then we will use function to extract the spans.

A screen shot of a computer program

AI-generated content may be incorrect.Example:

We see from this example if we had batch\_size = 2, sequence\_length=4, and fixed probabilities, we would have the indexes of head/tail from the probabilities.

We have similar function to extract triples (head, relation, tail) from the tail/head probabilities in the end of forward/backward extraction.

In general, the idea about the neural network is (I will talk forward extraction but the same for backward):

1. Create head start/end probabilities from passing token embeddings into fully-connected neural network and pass them into sigmoid activation

A black and white background with black text

AI-generated content may be incorrect.A white diamond with black text

AI-generated content may be incorrect.A black text on a white background

AI-generated content may be incorrect.

1. Extract subject set S representing a vector of the subjects

A green hexagon with white text

AI-generated content may be incorrect.Where each represents a subject by averaging the embeddings of start token and end token



1. We use predefined R which is relation embeddings got in the file prep\_relations where we save a tensor representing the relationships embeddings in BERT space (by taking last two hidden states of all tokens and use average poolings
2. A screenshot of a computer screen

   AI-generated content may be incorrect.For each subject extracted, we can extract the relations using semantic relation guidance, BRASK adopts specific attention mechanisms to compute attention scores that reflect a different role of each token. h\_g is pre-defined before representing the sentences in BERT space. attention scores that measure different role played by each token. represents fine-grained sentence representations.

A close-up of math equations

AI-generated content may be incorrect.

1. A screenshot of a computer

   AI-generated content may be incorrect.After we extract the objects. We do this by fusing: Subject representations and fine-grained sentence expressions of the semantic relations

A math equations with black letters

AI-generated content may be incorrect.

1. A diagram of a computer

   AI-generated content may be incorrect.We feed into fully connected neural network to obtain the start and end probabilities of each corresponding object

A black text on a white background

AI-generated content may be incorrect.

Backward triple extraction:

We do the same as forward extraction but:

* instead of using semantics relation guidance (BERT embeddings for relation), we use knowledge relation guidance (extract knowledge relations from the training of the TransE to our knowledge graph)
* Instead of extracting the subject and using guidance to extract objects, we extract objects and use knowledge to extract subjects

A yellow text with black letters

AI-generated content may be incorrect.

3. Experimental results

Provides an overview of the dataset used for experiments, the metrics used for evaluating performances, and the experimental methodology. Presents experimental results as plots and/or tables

A screen shot of a computer

AI-generated content may be incorrect.

I had 4815483 descriptions record before the cleaning, then I had 2786974 descriptions after cleaning

\*\*\*\*\*\*\*\*\*\* so in this case we can call it wikidata2.7m.\*\*\*\*\*\*\*delete\*\*\*\*

4. Concluding remarks

Provides a critical discussion on the experimental results and some ideas for future work