Combining Contextual Words and Knowledge Graph Embeddings

NLP M2

IDMC,University of Lorraine

Introduce The Team :

1. Fatima Habib (computer science)
2. Võ Tuấn Anh(Linguistics)
3. Minh Huong Ngo(Linguistics)
4. Asmaa Demny(Linguistics)

Aim of the project :

- In this project we compare the performance of Knowledge Graph embeddings vs Contextual words embedding vs The combination of the two techniques in applying many in many NLP Tasks like :
    a. Entity resolution :is the task of checking if two words or KG nodes represent the same entity.
    b. Textual entailment detection:is a task for which we determine if a chunk of text logically entails another.
- Follow one of the [1] future work directions :
    a. applying PCA on the embeddings in order to test for the “curse of dimensionality”
    b. using a simple, but higher-capacity, model such as MLP
- Use different Classifier : in [1] they use Logistic Classifier we are going to use Neural Networks instead .

Data sets :

Last year’s choice: **Freebase: Open collaborative KB**

Freebase 15K: reasonable number of entities

=> *for entity-focused and KG task*

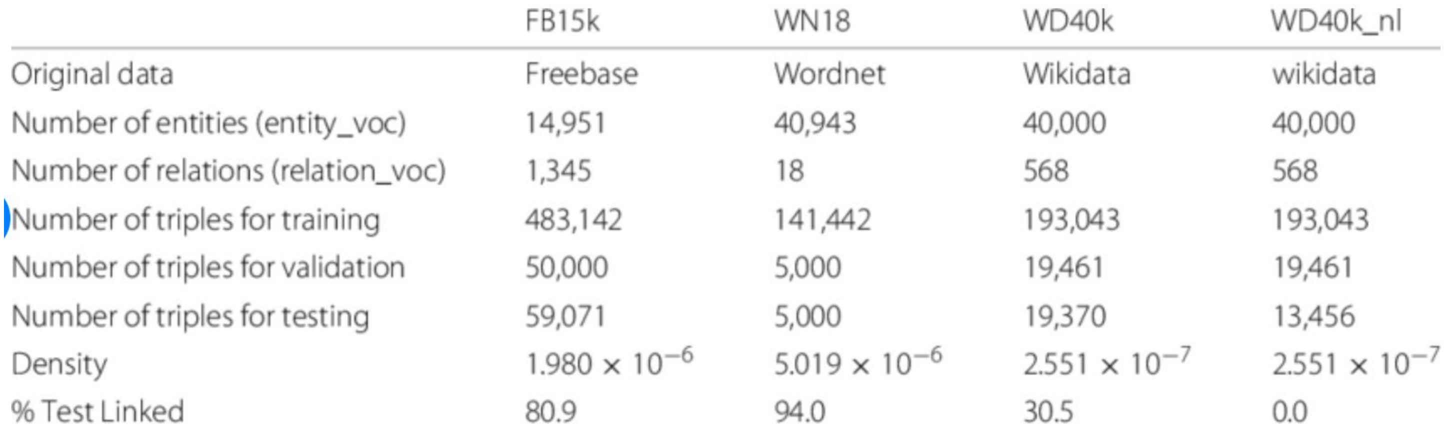
Freebase-NewYorkTimes: contain surface realization of triples

=> *for relation type prediction task*

=> *filter FB-NYT for relations and entities that were found in FB15k to do KG task*

Some possible Dataset choice

https://www.researchgate.net/figure/Details-of-FB15k-WN18-WD40k-and-WD40k-nl_tbl1_332831254



	FB15k	WN18	WD40k	WD40k_nl
Original data	Freebase	Wordnet	Wikidata	wikidata
Number of entities (entity_voc)	14,951	40,943	40,000	40,000
Number of relations (relation_voc)	1,345	18	568	568
Number of triples for training	483,142	141,442	193,043	193,043
Number of triples for validation	50,000	5,000	19,461	19,461
Number of triples for testing	59,071	5,000	19,370	13,456
Density	1.980×10^{-6}	5.019×10^{-6}	2.551×10^{-7}	2.551×10^{-7}
% Test Linked	80.9	94.0	30.5	0.0

Evaluations

- Extrinsic evaluation tasks:
  - Relation prediction
  - Entity classification
  - Entity resolution
  - Textual entailment detection
  - Triple (fact) classification
- Evaluation metrics: Precision@n, Mean Average Precision@k, Mean Reciprocal Rank

Tools and Framework:

- Python (spacy- NLTK -scikit learn -tensorflow)

- ELMo : Contextual Language Embedding
- ComplEXFramework

References :

1. Dieudonat, Léa & Han, Kelvin & Leavitt, Phyllicia & Marquer, Esteban. (2020). Exploring the Combination of Contextual Word Embeddings and Knowledge Graph Embeddings.
2. https://github.com/villmow/datasets_knowledge_embedding
3. https://github.com/nchah/freebase-triples

Aim of the project :

- In this project we compare the performance of Knowledge Graph embeddings vs Contextual words embedding vs The combination of the two techniques in applying many in many NLP Tasks like :
    a. Entity resolution :is the task of checking if two words or KG nodes represent the same entity.
    b. Textual entailment detection:is a task for which we determine if a chunk of text logically entails another.
- Follow one of the [1] future work directions :
    a. applying PCA on the embeddings in order to test for the “curse of dimensionality”
    b. using a simple, but higher-capacity, model such as MLP
- Use different Classifier : in [1] they use Logistic Classifier we are going to use Neural Networks instead .

Data sets :

Last year’s choice: **Freebase: Open collaborative KB**

Freebase 15K: reasonable number of entities

=> for entity-focused and KG task

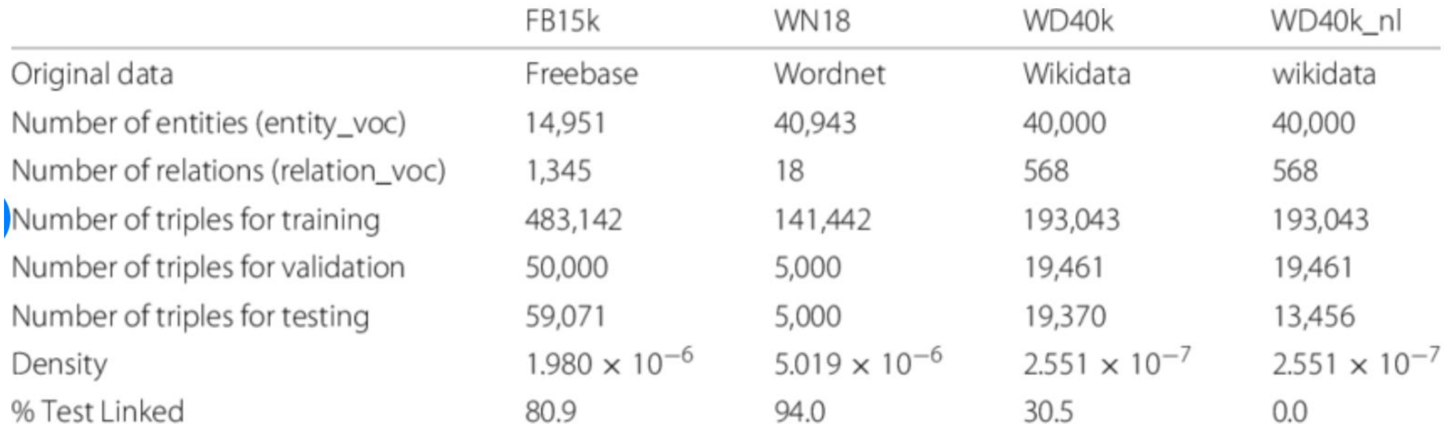
Freebase-NewYorkTimes: contain surface realization of triples

=> for relation type prediction task

=> filter FB-NYT for relations and entities that were found in FB15k to do KG task

Some possible Dataset choice

https://www.researchgate.net/figure/Details-of-FB15k-WN18-WD
1254



	FB15k	WN18	WD40k	WD40k_nl
Original data	Freebase	Wordnet	Wikidata	wikidata
Number of entities (entity_voc)	14,951	40,943	40,000	40,000
Number of relations (relation_voc)	1,345	18	568	568
Number of triples for training	483,142	141,442	193,043	193,043
Number of triples for validation	50,000	5,000	19,461	19,461
Number of triples for testing	59,071	5,000	19,370	13,456
Density	1.980×10^{-6}	5.019×10^{-6}	2.551×10^{-7}	2.551×10^{-7}
% Test Linked	80.9	94.0	30.5	0.0

Evaluations

- Extrinsic evaluation tasks:
  - Relation prediction
  - Entity classification
  - Entity resolution
  - Textual entailment detection
  - Triple (fact) classification
- Evaluation metrics: Precision@n, Mean Average Precision@k, Mean Reciprocal Rank

Tools and Framework:

- ❖ Python (spacy- NLTK -scikit learn -tensorflow)

- ❖ ELMo : Contextual Language Embedding
- ❖ ComplEXFramework

References :

1. Dieudonat, Léa & Han, Kelvin & Leavitt, Phyllicia & Marquer, Esteban. (2020). Exploring the Combination of Contextual Word Embeddings and Knowledge Graph Embeddings.
2. https://github.com/villmow/datasets_knowledge_embedding
3. https://github.com/nchah/freebase-triples

THANK YOU!