Cross-lingual ontology matching with CIDER-LM

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Javier Vela, Jorge Gracia

Introduction

CIDER-LM uses a pre-trained multilingual **language model** based on transformers [1], fine-tuned using the openly available portion of the MultiFarm dataset.

The model calculates the **vector embeddings** of the labels associated to every ontology entity and its context. The confidence degree between matching entities is computed as the **cosine similarity** between their associated embeddings.

CIDER-LM is novel in the use of multilingual language models for cross-lingual ontology matching.

Language Model

CIDER-LM relies on *distiluse-base-multilingual-cased-v2*, which is pre-trained on *Semantic Textual Similarity* and uses the SBERT architecture [2]. The model supports **more than 50 languages**.

The system uses a fine-tuned version of the model, specialized on the task of obtaining similarities between two entity labels in different languages. Given the labels and context from the training dataset ontologies, the model is trained on reducing the *CosineSimilarityLoss*.

Methods

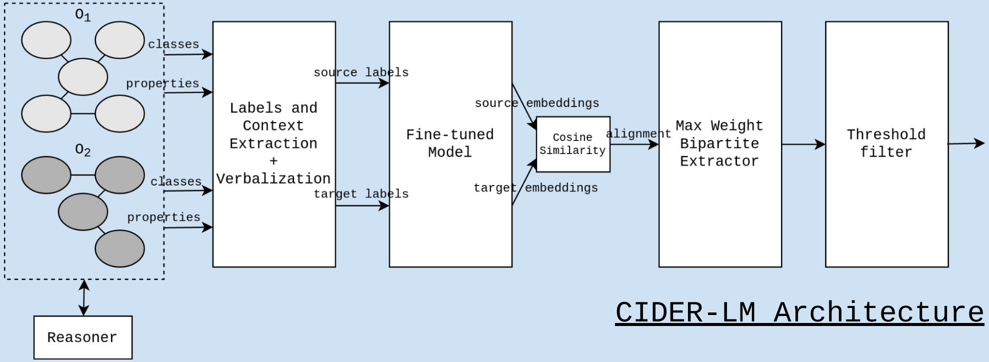
Verbalization
Entity labels are extended with their context; using parent and child (classes) and domain and range (properties)

Language Model
The model maps every entity to an embedding in a 512 dimensional dense vector space.

Cosine Similarity
The distance between the embeddings coming from the two different ontologies are calculated.

Hungarian Algorithm
The alignment is reduced to an 1 to 1 mapping.

Threshold
A threshold value of 0.5 is used to remove all correspondences with low confidence.



CIDER-LM Architecture

MultiFarm Results

Prec.	F-m.	Rec.	Time(Min)
0.16	0.25	0.58	157

In its first participation in the OAEI, CIDER-LM participated in the **MultiFarm** track only.

The obtained results in MultiFarm are intermediate in terms of F-measure (3rd of 6), but very good in terms of **recall** (the best result of any MultiFarm OAEI edition).

Conclusion

CIDER-LM has potential to improve in several ways:

- Including more context **features** to obtain more representative embeddings representations for ontology entities.
- Adjusting the **threshold** value to balance the precision and recall results.
- Involving more sophisticated techniques on the **fine-tuning** of the model can provide a more general model that behaves better with ontologies different from the ones seen in training.

References

[1] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Attention Is All You Need*, in: Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017
[2] N. Reimers, I. Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT Networks*, in: Proc. of EMNLP-IJCNLP 2019, Association for Computational Linguistics, 2019, pp. 3982–3992