Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA18209

Grantee name: E-COST-GRANT-CA18209-35d4db63

Details of the STSM

Title: Relation Acquisition from Large Language Models (LLMs): Approaches and Evaluation

Start and end date: 11/09/2023 to 30/09/2023

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

(max. 500 words)

During the first week of the STSM, the grantee participated in the 4th Conference on Language, Data and Knowledge (LDK 2023), in the host city, Vienna, organised by the grant host. He presented two posters, one of which on Relation Acquisition in Portuguese from the Large Language Model GPT-3, thus related to the topic of the STSM. The poster and the participation in general enabled networking and fuelled interesting discussions on the topic of the STSM. Specifically, in the workshop "Deep Learning, Relation Extraction and Linguistic Data with a Case Study on BATS" (http://dl4ld2023.mruni.eu/), the translation of the BATS dataset [1] to some languages, its validation and utilisation, were discussed. Most of the experiments performed in the remaining two weeks were supported by this dataset, recently translated to several languages in the scope of Nexus Linguarum.

Another important outcome was the proposal of a new workshop on Deep Learning and Linked Data, to be held, if accepted, as a satellite of the LREC-COLING 2024 conference (https://lrec-coling-2024.org/).

The remaining two weeks were focused on the main goal of the STSM: the collaboration in Relation Acquisition from Large Language Models in a multilingual context.

This was tackled with experiments in the lexico-semantic part of the BATS dataset, mainly performed with multilingual masked language models like the XLM-RoBERTa-base multilingual [2]. Those included zero and five-shot Analogy Completion in a set of languages, and a variation that we called Analogy-based Translation, which takes advantage of the parallel nature of the dataset. Both tasks were

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

approached as masked language modelling and as multiple-choice question answering. The latter, explored as a way to deal with multi-token words, resorted to the FitBERT library (https://github.com/writerai/fitbert).

The grantee also engaged in discussions regarding additional experiments on the same topic, performed by other grantee (Lucía Pitarch), namely on Multilingual Relation Classification, including language transfer visualisations; and on an RDF representation for the dataset, where, after considering other vocabularies, the Cross-Linguistic Data Formats (https://cldf.clld.org/) were explored. This will enable to link the dataset to other linguistic resources.

Towards the publication of the BATS translation and its utilisation in experiments, scripts were developed for pre-processing. This helped in the computation of simple figures and on the identification of annotation issues and missing information. Some of the previous were fixed directly, while others required the language authors to be contacted.

A scientific paper describing the new dataset and the previous experiments started to be drafted, having in mind its submission to LREC-COLING 2024. A second paper was considered for accommodating all the possible experiments, in a venue to be decided in the future.

[1] Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In Proceedings of NAACL 2016 Student Research Workshop, pages 8–15. ACL.

[2] Conneau, A., Khandelwal, K., Goyal, et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451).

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

(max. 500 words)

The following goals were achieved with the STSM:

- Preparation of the Multilingual Lexical BATS dataset (baptised as MultiLex BATS) for its publication.
- First results on Multilingual Relation Classification in the MultiLex BATS.
- First visualisations of language transfer in the previous task.
- First results on Analogy Completion in several languages covered by MultiLex BATS.
- Proposal of a new task for assessing language transfer, called Analogy-Based translation, which can be assessed in MultiLex BATS.
- Proposal of an RDF representation of MultiLex BATS.
- Preparation of a workshop proposal on Deep Learning and Linked Data (DLnLD) for LREC-COLING 2024.

Some of the previous will be described in the paper to be submitted to LREC-COLING, which we started to write during the STSM. A second paper, focused on some of the experiments, is also planned.

The previous goals contribute to tasks in several Working Groups of the Nexus Linguarum COST action, namely:

- WG1: the representation of MultiLex BATS in RDF, and the concern on covering as many languages as possible, also including under-resourced languages, are related to T1.1 LLOD

modelling; T1.3 Cross-lingual data interlinking, access and retrieval in LLOD; and T1.5 Development of the LLOD cloud for under-resourced languages and domains.

- WG2: the experiments focused on multilingual Knowledge Acquisition are related to T2.1 LLOD in Knowledge extraction and 2.2 LLOD in Machine Translation.
- WG3: all experiments resorted to neural language models and are thus related to T3.2 Deep Learning and Neural Approaches for Linguistic Data. The workshop is also related to this task.
- WG4: the initiative of creating MultiLex BATS is part of the Use Case 4.1.3 Acquiring RDF Relations with Neural Language Models.