Boosting MultiFarm Track with Turkish Dataset

Abderrahmane Khiat[1], Beyza Yaman[2], Giovanna Guerrini[2], Ernesto Jiménez-Ruiz[3], and Naouel Karam[1]

[1] Human-Centered Computing Lab, Freie Universität Berlin, Germany
[2] DIBRIS, University of Genoa, Italy
[3] University of Oslo, Norway

1 Introduction

The evolution of semantic structured data, such as those behind the deep web or social networks, requires mapping between sources to enable a high level integration. Several ontology matching systems have been developed to establish mappings between multilingual ontologies, however, employing these systems in real world requires an assessment of the ontologies capability and performance which is conducted by the MultiFarm Track. Yet, this track still lacks of ontologies from different language families. In this paper, we contribute to the OAEI initiative with a Turkish dataset to extend the coverage of languages for the matching systems.

2 The Proposed Dataset

The Turkish language comes from a different branch of the language family than the existing datasets in the Multifarm Track, we believe it will add a different perspective to the assessments of matching systems. The dataset is valuable for the OM domain for several reasons: *i)* Since the Mutltifarm track is composed of a set of ontologies of the conference domain, to the best of our knowledge no such dataset exists for the conference domain in Turkish, *ii)* we will integrate Turkish datasets to the OAEI campaign to assess the performance of cross-lingual ontology alignment systems along with other languages and *iii)* we will close the gap of lacking datasets for Turkish in the OAEI.

We followed the steps detailed in [3] to create our dataset, to validate it and then to generate the reference alignments for other ontologies. The Multifarm track has been translated from English to Turkish semi-automatically, and then reviewed and corrected by a professional English-Turkish speaker. During dataset generation, we have taken advantage of our experiences of generating Arabic datasets [2]. We first generated the Turkish ontologies via regular expressions with the regex API and, then, translated entities are replaced by the original ones. Finally, the alignments between Turkish and other languages are constructed by replacing English entity IDs by Turkish entity IDs.

However, we have a different level of difficulty for generating alignments between English-Turkish than we have experienced when we were constructing the alignments for Arabic datasets. The difficulty for English-Arabic datasets was to find an automatic solution for generating alignments between files where each one contains a high number of semantic correspondences. On the other hand, thanks to our automation of the framework, it was easier to create English-Turkish alignments by implementing a generator

for the solution. The solution consists of (1) considering the dataset that contains a high number of semantic correspondences (e.g. English-French), (2) replacing English entities by the new language (e.g. Turkish) (3) replacing entities of the other language (e.g. French) with corresponding English entities, using the alignments of the same ontologies between English and the other language (e.g. French in this case).

3 Experiments

The experimental study conducted on the Turkish datasets is performed using the CroLOM system due to its good results (ranked third) obtained in the OAEI2016 edition. CroLOM[1] uses the Yandex translator, NLP techniques and a similarity computation based on the categories of words and synonyms. The experimental results are presented in Table 1 for each language pair. The results are good for the pairs English and Spanish. However, they are less satisfactory for the pairs Chinese, Arabic and German. This is explained by the fact that CroLOM uses English as pivot to align multilingual ontologies. We can also observe that, on average Turkish ontologies bring an additional complexity to the Multifarm track, w.r.t. the results without Turkish dataset obtained via CroLOM.

Table 1: Results of the CroLOM System on the Turkish Dataset

Dataset Pairs	H-Mean Pre.	H-Mean F-mea.	H-Mean Rec.
Arabic-Turkish	0.77	0.29	0.18
English-Turkish	0.74	0.47	0.34
German-Turkish	0.59	0.36	0.26
Czech-Turkish	0.71	0.40	0.27
Chinese-Turkish	0.47	0.25	0.17
Spanish-Turkish	0.65	0.48	0.38
French-Turkish	0.60	0.42	0.33
Dutch-Turkish	0.64	0.43	0.33
Portuguese-Turkish	0.70	0.45	0.34
Russian-Turkish	0.69	0.41	0.29

The CroLOM system completed all the tests involving the Turkish language and the experimental study shows that the dataset is suitable to evaluate state-of-the-art ontology matching systems.

Acknowledgements We would like to thank Lecturer Nuriye In for her contributions to the corrections of the datasets.

References

1. A. Khiat. Crolom: cross-lingual ontology matching system results for OAEI 2016. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Japan*, pages 146–152, 2016.
2. A. Khiat, G. Diallo, B. Yaman, E. Jiménez-Ruiz, and M. Benaissa. Abom and adom: Arabic datasets for the ontology alignment evaluation campaign. In *ODBASE 2015*, pages 545–553.
3. C. Meilicke, R. Garcia-Castro, F. Freitas, W. R. van Hage, E. Montiel-Ponsoda, R. R. de Azevedo, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, A. Tamilin, C. T. dos Santos, and S. Wang. Multifarm: A benchmark for multilingual ontology matching. *J. Web Sem.*, 15:62–68, 2012.