ADOM: Arabic Dataset for Evaluating Arabic and Cross-lingual Ontology Alignment Systems

Abderrahmane Khiat[1], Moussa Benaissa[1], and Ernesto Jiménez-Ruiz[2]

[1]LITIO Laboratory, University of Oran1 Ahmed Ben Bella, Oran, Algeria
[2]Department of Computer Science, University of Oxford, United Kingdom

Abstract. In this paper, we present ADOM, a dataset in Arabic language describing the conference domain. This dataset was created for two purposes (1) analysis of the behavior of matchers specially designed for Arabic language, (2) integration with the multifarm dataset of the Ontology Alignment Evaluation Initiative (OAEI). The multifarm track evaluates the ability of matching systems to deal with ontologies described in different natural languages. We have tested the ADOM dataset with the LogMap ontology matching system. The experiment shows that the ADOM dataset works correctly for the task of evaluating cross multilingual ontology alignment systems.

1 Introduction

Ontology alignment is defined as the identification process of semantic correspondences between entities of different ontologies in order to ensure the semantic interoperability [1]. However, the automatic identification of correspondences between ontologies is very difficult due to (a) their conceptual divergence [8], and (b) to the use of different naming conventions or languages. In the literature there are several systems that deal with the (semi) automatic alignment of ontologies [1, 12, 11]. These systems are (typically) primarily based on the lexical similarity of the entity labels. Matching ontologies in different languages is challenging due to misinterpretations during the translation process. Ontologies in Arabic language brings even more challenges due to special features of the language. Among the reasons that make ontology alignment in Arabic language very difficult we can quote [6]:

1. The Arabic script (no short vowels and no capitalization).
2. Explosion of ambiguity (in average 2.3 per word in other languages to 19.2 in Arabic) by Buckwalter (2004) [5].
3. Complex word structure, for example the sentence ورأيتهم can be translated in English language as and I saw them.
4. The problem of Normalization, for example آ ، إ ، أ ، ا → ا i.e. losing distinction آن ، إن ، أن
5. The Arabic language is one of the pro-drop languages, i.e. languages that allow speakers to omit certain classes of pronouns

Table 1: Top systems in the multifarm track

OAEI	Top Systems	Precision	F-measure	Recall
2012	YAM++	0.50	0.40	0.36
2013	YAM++	0.51	0.40	0.36
2014	AML	0.57	0.54	0.53
2014	LogMap	0.80	0.40	0.28
2014	XMap	0.31	0.35	0.43

In this paper, we present ADOM, a dataset in Arabic language describing the conference domain. We have created this dataset by translating and improving all ontologies of the conference track [13] of the OAEI campaign. We summarize below the objectives of the developed dataset: (1) Analysis and evaluation of the behaviour of matchers designed for Arabic language. Here, the real questions are: (a) could the state of the art systems handle efficiently the ontologies described in Arabic language? (b) Are external knowledge resources for Arabic language available such as WordNet? (2) Integration with the multifarm track [14] of the OAEI campaign.[1] The multifarm track evaluates the ability of matching systems to deal with ontologies described in different natural languages. The question here, concerns to the performance of the translator used to align multilingual ontologies?

The rest of the paper is organized as follows. First, in Section 2, we discuss the top systems that participated in the last editions of the multifarm track. In section 3 we describe the ADOM dataset. Section 4 contains the experiment results. Finally, some concluding remarks and future work are presented in Section 5.

2 Related Work

In this section we discuss the main ontology matching systems that have participated in the multifarm track. Most of such systems use a translation tool to deal with the cross-lingual ontology alignment. The XMap system [2] uses an automatic translation for obtaining correct matching pairs in multilingual ontology matching. The translation is done by querying Microsoft Translator for the full name. The AML system [4] uses an automatic translation module based on Microsoft Translator. The translation is done by querying Microsoft Translator for the full name (rather than word-by-word). To improve performance, AML stores locally all translation results in dictionary files, and queries the Translator only when no stored translation is found. The LogMap system [10] that participated in the OAEI 2014 campaign used a multilingual module based on Google translate [3]; however the new version of the LogMap system uses both Microsoft and Google translator APIs [9]. The YAM++ system [7] uses a multilingual translator based on Microsoft Bing to translate the annotations to English. Table 1 summarizes the results of the top systems in the multifarm track.

[1] ADOM has already been integrated within the OAEI 2015 multifarm dataset: `http://oaei.ontologymatching.org/2015/multifarm/index.html`

3 The ADOM Dataset

The dataset is constituted of seven ontologies in Arabic language. These ontologies describe the conference domain and are based on the ontologies of the OAEI conference track [13]. We justify the proposal of our dataset by the following points: (1) The OAEI campaign, which is the most known evaluation campaign for testing the performance of ontology matching systems, lacked a test case involving ontologies in Arabic language. (2) To the best of our knowledge, no such dataset exists yet in Arabic language.[2] (3) Furthermore, there are several contexts such as Web information retrieval where the ontology matching systems are needed both in inter-multilingual ontologies and intra-Arabic ontologies.

We have developed our dataset relying on the conference and multifarm tracks of the OAEI. In order to develop the Arabic ontologies and reference alignments for the ADOM dataset we proceeded as follows.

3.1 Step 1: Translation of Ontology Entities

In this step, we have identified the concepts, object and data-type properties of the ontologies, for example we can list the concept ”البحث (paper)”, data-type property ”لديه اسم (has name) and object property ”لديه موقع على رابط (has website at URL), etc. We have semi-automatically translated the ontologies in English and French by considering the context of the ontologies (i.e., the conference domain). For example, if we translate simply the concept ”paper” we get ”ورقة” in Arabic language but ”ورقة” is not the correct concept if we consider the context of conference and some information from conference websites in Arabic language. Then the correct concept of ”paper” becomes ”البحث”.

3.2 Step 2: Generation of Reference Alignments

We have reused the available reference alignments among the ontologies in the multifarm track to generate the new reference alignments for ADOM. For example, in the reference alignment for ontologies in Arabic language, we can list the concept ”الحدث (event)” of the ontology Confof is equivalent to the concept ”نشاط (activity)” of the ontology Iasted. In the reference alignment for ontologies in Arabic and French languages, we can list the concept ”éditeur (Editor)” of the ontology conference is equivalent to the concept ”المحرر” of the ontology Cmt.

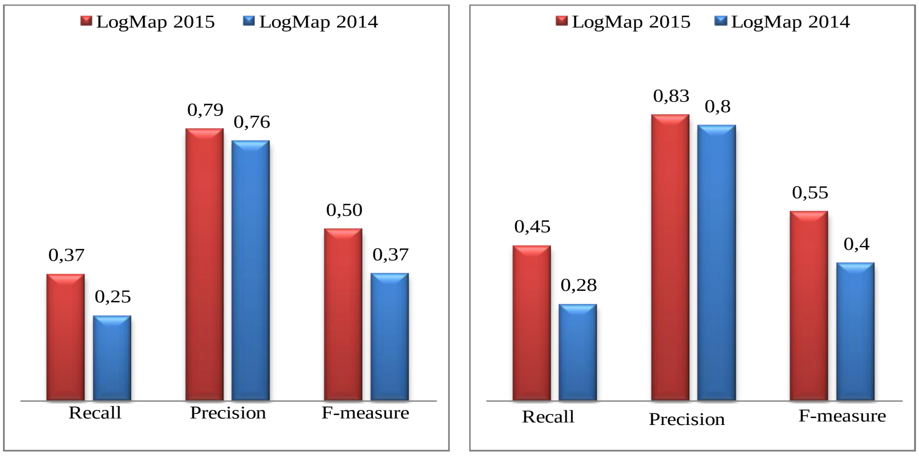
3.3 Step 3: Validation by a Linguistic Expert

Our dataset was validated by a linguistic expert with regard to the translation of concepts and properties. Furthermore we also checked the correctness of the new reference alignments.

[2] Note that, in the literature one can find datasets in Arabic language applied to other domains different from Ontology Matching (e.g. [15, 16])

4 Experimental Study

In order to evaluate the ADOM dataset, we have used the LogMap system which is one of the top ontology alignment systems on multifarm track (see Table 1). The purpose of this evaluation is to show that the ADOM dataset is suitable to test ontology matching systems that implement multilingual support.



(a) LogMap results on ADOM dataset

(b) LogMap System on Multifarm Track

Fig. 1: LogMap 2014 and 2015 results on ADOM and in all multifarm dataset.

We have tested the ADOM dataset with two versions of LogMap system. The first version, which has participated in the OAEI 2014, uses the Google translator API. The second, which aims at participating in the OAEI 2015, uses both the Microsoft and Google translators APIs. Figure 1 summarizes the average results, in terms of precision, recall and F-measure, obtained by LogMap on the ADOM dataset (Fig. 1a) and on all multifarm tests (Fig. 1b). We can appreciate that, on average, the ADOM dataset brings an additional complexity to the multifarm track, with regard the results obtained by LogMap. Note that, we aim at obtaining a more comprehensive evaluation during the OAEI 2015 evaluation campaign to confirm this fact.

5 Conclusion

In this paper we have presented ADOM, a new dataset in Arabic language describing the conference domain. This dataset has been created for two purposes 1) studying and developing specific ontology alignment methods to align ontologies in Arabic language, 2) evaluating the ability of state of the art ontology matching systems to deal with ontologies in Arabic. The experimental study shows that ADOM dataset is suitable in practice. Furthermore, ADOM has already been integrated within the multifarm dataset and it will be evaluated in the OAEI 2015 campaign.

As future challenges, we aim at (1) developing a large corpus of ontologies and dictionaries for the Arabic language, (2) adapting state of the art NLP tools to align ontologies in Arabic language, (3) improving the state of the art translators dedicated to the Arabic language.

References

1. J. Euzenat and P. Shvaiko, “Ontology Matching”, Springer-Verlag, Heidelberg, 2013.
2. W. Djeddi and M. T.Khadir, “XMap++ results for OAEI 2014”. In Proceedings of the 9th International Workshop on Ontology Matching ISWC 2014, pp. 163169, Italy, 2014.
3. E. Jiménez-Ruiz, B. C. Grau, W Xia, A. Solimando, X. Chen, V. Cross, Y. Gong, S. Zhang and A. Chennai-Thiagarajan, “LogMap family results for OAEI 2014”. In Proceedings of the 9th Workshop on Ontology Matching ISWC 2014, pp. 126-134, Italy, 2014.
4. D. Faria, C. Martins, A. Nanavaty, A. Taheri, C. Pesquita, E. Santos, I. F. Cruz and F. M. Couto, “AgreementMakerLight results for OAEI 2014”. In Proceedings of the 9th Workshop on Ontology Matching ISWC 2014, pp. 105-112, Italy, 2014.
5. T. Buckwalter, “Arabic Morphological Analyzer Version 2.0”. LDC catalog number LDC2004L02, 2004.
6. A. Farghaly, “Arabic NLP: Overview, state of the art, challenges and opportunities”, In The International Arab Conference on Information Technology, ACIT2008, Tunisia, 2008.
7. D. Ngo and Z. Bellahsene, “YAM++ results for OAEI 2013”, In Proceedings of the 8th Workshop on Ontology Matching ISWC 2013, pp. 211-218, Australia, 2013.
8. P. Bouquet, J. Euzenat, E. Franconi, L. Serafini, G. Stamou and S. Tessaris “Specification of a Common Framework for Characterizing Alignment”, Deliverable 2.2.1, Knowledge Web NoE, Technical Report, Italy, 2004.
9. E. Jiménez-Ruiz et al. “LogMap family results for OAEI 2015”. In Proceedings of the 10th Workshop on Ontology Matching ISWC 2015, pp., USA, 2015.
10. E. Jiménez-Ruiz, Bernardo Cuenca Grau, Yujiao Zhou and Ian Horrocks. “Large-Scale Interactive Ontology Matching: Algorithms and Implementation”. In: ECAI. 2012.
11. Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. O. Kempf, P. Lambrix, S. Montanelli, H. Paulheim, D. Ritze, P. Shvaiko, A. Solimando, C. Trojahn-dos-Santos, O. Zamazal and B. Cuenca Grau, “Results of the Ontology Alignment Evaluation Initiative 2014”, 9th Workshop on Ontology Matching, 2014.
12. B. Cuenca Grau, Z. Dragisic, K. Eckert, J. Euzenat, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. Oskar Kempf, P. Lambrix, A. Nikolov, H. Paulheim, D. Ritze, F. Scharffe, P. Shvaiko, C. Trojahn dos Santos, O. Zamazal, “Results of the Ontology Alignment Evaluation Initiative 2013”. 8th Workshop on Ontology Matching, 2013.
13. O. Svab, V. Svatek, P. Berka, D. Rak and P. Tomasek, “OntoFarm: Towards an Experimental Collection of Parallel Ontologies”, In: Poster Track of ISWC 2005, Galway, 2005.
14. C. Meilicke, R. Garca-Castro, F. Freitas, WR. Van Hage, E. Montiel-Ponsoda, R.R. De Azevedo, H. Stuckenschmidt, O. vb-Zamazal, V. Svtek and A. Tamilin, “MultiFarm: A benchmark for multilingual ontology matching”. Web Semant. Sci. Serv. Agents World Wide Web. Vol. 15, pp. 6268, 2012.
15. I. Bounhas, B. Elayeb, F. Evrard, and Y. Slimani, ”ArabOnto: Experimenting a new distributional approach for Building Arabic Ontological Resources”. In International Journal of Metadata, Semantics and Ontologies, Inder-science, Vol. 6, No. 2, pp. 81-95, 2011.
16. O. Ben Khiroun, R. Ayed, B. Elayeb, I. Bounhas, N. Bellamine Ben Saoud and F. Evrard, “Towards a New Standard Arabic Test Collection for Mono- and Cross-Language Information Retrieval”, In the Proceedings of 19th International Conference on Application of Natural Language to Information Systems (NLDB), 2014.