SimCat Results for OAEI 2016

Abderrahmane Khiat[1], Elhabib Abdelillah Ouhiba[2], Mohammed Amine Belfedhal[3] and Chihab Eddine Zoua[4]

[1]LITIO Laboratory, University of Oran1 Ahmed Ben Bella, Oran, Algeria
[2]LAMOSI Laboratory, Oran University of Science and Technology - Mohamed Boudiaf
[3]EEEDIS Lab, University Djillali Liabes, Sidi Bel-Abbes, Algeria
[4]Ooredoo Algiers, Algeria
Email: abderrahmane_khiat@yahoo.com, ouhiba.ab@gmail.com,
Mohammed.belfedhal@gmail.com, z.chiheb.e@gmail.com

Abstract. Recently, the multilingualism issue has attracted considerable attention in the ontology matching field. Designed for this purpose, the SimCat system uses the Yandex translator and similarity computation based on the categories of the words. This is the first participation of SimCat in OAEI 2016 evaluation campaign and the obtained results are quite promising.

1 Presentation of the System

The Semantic Web relies on ontologies to describe the content of different information sources in order to overcome the heterogeneity issue and achieve their semantic interoperability [12, 14]. However, these ontologies are heterogeneous, distributed and even they are described in different languages. A solution to this heterogeneity is to use ontology alignment to bridge the semantic gap between these ontologies [11]. The ontology alignment system receives as input two or more ontologies and generates as output a set of semantic correspondences between the entities of the ontologies that are being processed [3, 2]. Indeed, these semantic correspondences are the bridges that hold the heterogeneous ontologies together and ensure their semantic interoperability. Moreover, with the enormous volume of ontologies already available on the web and their constant evolution, manual identification of semantic correspondences is not feasible [14]. Therefore, ontology alignment tools are required to have the ability of identifying semantic correspondences between entities of different ontologies in an automated way. However, the automatic identification of semantic correspondences is not a trivial task due to the conceptual diversity between the ontologies [4].

Performing an automatic ontology alignment task between mono-language ontologies such as English is difficult, however, the task is even more challenging when it comes to multilingual ontologies. Most existing approaches implement a direct strategy[15] i.e. using machine translation. However, the matching task is challenging for these approaches due to misinterpretations during the translation process.

The research conducted on direct strategy leaves many questions to address such as (1) is the use of various translators has a different impact on the output of the translation? (2) is the translation into a pivot language (English) performing better output than

a translation from language to another? and (3) how to proceed when translators give poor results?

The multifarm[10] track has been integrated in the Ontology Alignment Evaluation Initiative (OAEI) in 2012 with the goal of estimating and comparing different techniques and systems related to multilingual ontology alignment. From 2012 to 2014 the multifarm track contains conference ontologies[9] described in eight different languages (i.e., Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish). However, in 2015 the multifarm includes the Arabic language [13, 14].

Back to results of the systems involved in previous editions (from 2012 to 2015) [5–8] of multifarm track, we have observed that the best system (in all previous OAEI editions) achieved an F-measure of 0.51 [15]. This is surprising, in spite of many research works that have been established in the field of multilingual ontology matching.

The proposed system also implements a direct strategy and its aim is to highlight the translator used and similarity calculated using the categories of the word.

1.1 State, Purpose, General Statement

In this paper, we describe our SimCat software, yet another cross-lingual ontology matching system. Unlike existing approaches which use well-known translators, SimCat employs the Yandex translatorr[1]. In addition, SimCat computes the similarities between translated entities based on the categories of the words.

1.2 Specific Techniques Used

The process of our system consists in the following successive steps.

Step 1: Extraction and Normalization In this step, our system extracts the entities of two ontologies to align. Then, it uses a segmentation technique to split labels into words; Finally, it converts all words in lower case.

Step 2: Translation and Cleaning In this step, SimCat translates the normalized entities using the Yandex translator into English as a pivot language. To the best of our knowledge, the Yandex translator has not been used before by multilingual ontology matching system. Our choice of Yandex translator is justified by the fact that it is one of largest search engine in the world and the obtained results are quite promising. However, we have used the English as a pivot language because the categories of the words which are used for similarity computation are in English language.

Once the translation is is carried out, SimCat employs NLP techniques. First, it eliminates the stop-words from translated entities; then it employs lemmatization and stemming. This step is necessary since the categories of the words are in that lemma form.

[1] `https://translate.yandex.com/?lang=es-en&text=administrar&ncrnd=5317`

Step 3: Similarity Computation In this step, our system computes the similarity between entities using the categories of words. This matcher is based on an open project named "Calculate Semantic Similarity".

The project[2] calculates the similarities between sentences and the results are stable. The description of the project is as follows: First, the list of words was obtained from using EOWL, then the categories for each word were calculated using the DISCO's semantics[3]. The semantic categories are obtained from disco as follows: (1) en-BNC-20080721 within 119 million tokens; (2) en-PubMedOA-20070501 within 181 million tokens and (3) en-wikipedia-20080101 within 267 million tokens. The matcher enhances the Vector-Space by the analysis found withing the Classifier4j, which does not take into account the semantic meanings of the words.

However, we have adapted it for our case. We have reprogram the matcher in a way that it can return the similarity value between words. We have some tests on the adapted matcher and the results are quite good.

Step 4: Identification of Alignment In this step, SimCat applies applies a filter to select candidate correspondences which possess the maximum similarity value in each line of Cartesian product between entities. Then it applies a second a filter to identify the correspondences that possess similarity value upper than a given threshold.

1.3 Adaptations Made for the Evaluation

We do not have made any specific adaptation for OAEI 2016 evaluation campaign regarding our SimCat system. All parameters are the same for aligning different ontologies of multifarm track.

1.4 Link to the set of provided alignments (in align format)

The result of SimCat system can be downloaded from OAEI 2016 website `http://oaei.ontologymatching.org/2016/results/multifarm/index.html`

2 Results

The SimCat system is yet another multilingual ontology alignment system. Designed for this purpose, we present the results obtained by running our SimCat system on multifarm tracks of OAEI 2016 evaluation campaign following website `http://oaei.ontologymatching.org/2016/results/multifarm/index.html`.

The multifarm track is constituted of seven ontologies. These ontologies describe the conference domain and are based on the ontologies of the OAEI conference track. These ontologies have been translated in nine different languages (since 2015 the Arabic language is included, Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish) and the corresponding alignments between these ontologies. The purpose of multifarm is to evaluate and compare the performance of matching approaches with a special focus on multilingualism.

[2] `http://wordnet.princeton.edu/`

[3] `www.linguatools.de/disco/disco_en.html`

3 General Comments

The evaluation conducted on SimCat system confirmed the following points:

- The results obtained from the Yandex translator API are quite promising.
- The similarity based on the categories of the words could provide good results.
- In overall, the SimCat system provides promising results by achieving a good F-Measure, however, it consumes 24 min as computation time for each task. This is considered as a drawback of the proposed system, since the multifarm contains 55 tasks.

4 Conclusion

In this paper, We have presented SimCat, an automatic matching system developed specifically for aligning multilingual ontologies. The SimCat system implements a matcher based on the categories of the words and a translation based on Yandex engine to find the semantic correspondences between different concepts of the two ontologies described in different natural languages. Regarding the first participation of SimCat system in OAEI2016, the results are acceptable, however there is much work to do in order to improve our system.

References

1. M. Ehrig, "Ontology Alignment: Bridging the Semantic Gap", Springer, 2007.
2. P. Shvaiko and J. Euzenat, "Ontology Matching: State of the Art and Future Challenges", IEEE Transactions on Knowledge and Data Engineering vol. 25 no. 1, pp. 158-176, 2013.
3. J. Euzenat and P. Shvaiko, "Ontology Matching", Springer-Verlag, Heidelberg, 2013.
4. 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.
5. M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, R. Granada, V. Ivanova, E. Jiménez-Ruiz, P. Lambrix, S. Montanelli, C. Pesquita, T. Saveta, P. Shvaiko, A. Solimando, C. Trojahn and O. Zamazal, "Results of the Ontology Alignment Evaluation Initiative 2015", 10th Workshop on Ontology Matching, 2015.
6. 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.
7. 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.
8. J. Aguirre, K. Eckert, J. Euzenat, A. Ferrara, W.R.v. Hage, L. Hollink, Ch. Meilicke, A. Nikolov, D. Ritze, F. Scharffe, P. Shvaiko, O. Svb-Zamazal, C. Trojahn, E. Jiménez-Ruiz, B. Cuenca-Grau and B. Zapilko:, "Results of the Ontology Alignment Evaluation Initiative 2012", 7th Workshop on Ontology Matching, 2012.

9. 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.
10. 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.
11. A. Khiat and M. Benaissa, A New Instance-Based Approach for Ontology Alignment. International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 11, No. 3, ISSN 1683-3198, 2015.
12. A. Khiat and M. Benaissa, Boosting Reasoning-Based Approach by Structural Metrics for Ontology Alignment. The Journal of Information Processing Systems (JIPS), 2015.
13. A. Khiat and M. Benaissa and Ernesto Jimnez-Ruiz ADOM: arabic dataset for evaluating arabic and cross-lingual ontology alignment systems. In Proceedings of the 10th International Workshop on Ontology Matching co-located with the 14th International Semantic Web Conference (ISWC 2015), USA, 2015.
14. A. Khiat, G. Diallo, B. Yaman, E. Jimnez-Ruiz and M. Benaissa, ABOM and ADOM: Arabic Datasets for the Ontology Alignment Evaluation Campaign. In Proceedings of the 14th International Conference (ODBASE 2015), Greece, 2015.
15. A. Khiat, CroLOM: Cross-Lingual Ontology Matching System Results for OAEI 2016. In Proceedings of the 12th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Japan, 2016.