CroLOM results for OAEI 2017

Summary of Cross-Lingual Ontology matching Systems at OAEI

Abderrahmane Khiat

Human-Centered Computing Lab, Freie Universität Berlin, Germany
abderrahmane.khiat@fu-berlin.de, abderrahmane_khiat@yahoo.com

Abstract. This paper presents the results obtained in the OAEI 2017 campaign by our ontology matching system CroLOM. CroLOM is an automatic system especially designed for aligning multilingual ontologies. This is our second participation with CroLOM in the OAEI and the results have so far been positive.

Keywords: Cross lingual Alignment, Multilingual Ontologies Survey, Ontology Matching, Yandex, Semantic Similarity, OAEI, Direct matching.

1 Introduction

With the growing number of ontologies defined in different languages, multilingualism has become an issue of major interest in ontology matching field. Multilingual ontology alignment, defined as the process of identification of semantic correspondences between entities of different ontologies described in different natural language, represents the solution to the problem of semantic interoperability between different sources of distributed information [1, 2]. Several methods have been elaborated to semantically align multilingual ontologies. These methods can be generally split into two main categories direct and indirect matching approaches [3]. The approaches of the first category are based on external resources (i.e. translation) to align cross-lingual ontologies. However, the approaches of the second category are based on the composition of alignments such as the work proposed in [4] where the authors reuse the mappings between ontologies that already exist.

In this study, we consider the approaches of the first category, since we develop an approach which implements a direct strategy. However, there are many questions regarding this approach to address the multilingualism issue. These questions are as follows: (1) Which machine translation should be used, (2) which translation path should be considered and (3) which ontologies features and dictionaries can be exploited. In the following paragraphs, we describe the points mentioned above.

First, several translators have been developed to translate automatically the text from one natural language to another. We can mention for example: Google, Bing, SDL and Gengo translators. Each translator has its specific characteristics such as: number of source/target languages and execution time. However, selecting one or several translators (by combining them) remains an open problem. This choice is crucial in ”direct approaches”, since they apply a monolingual matching techniques in cross-lingual ontology mapping.

Second, the translation path also plays an important role to resolve the heterogeneity problem. Two translation paths can be considered, *(i)* either considering the translation path from one to another or *(ii)* selecting a pivot language which is often the English language. This choice highly depends on available sources (dictionaries, thesaurus, etc.) in different natural languages. Most matching systems consider the translation path using English as a pivot language due to available sources in English language.

Finally, in some cases, the results of a translation machine could be poor, however, to avoid this situation some ontology features can be exploited such Description Logics.

Most matching systems which implement a direct translation approach uses a well-known translators mentioned above. The current work uses also a direct matching approach. However, unlike existing approaches, it addresses the multilingualism challenge by using *(a)* the Yandex translator[1], *(b)* a translation into a pivot language after applying NLP and *(c)* a similarity computation based on the categories of the words and synonyms.

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 CroLOM system. 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 continue our previous work [5, 20] by covering the main cross-lingual ontology matching systems that have participated in the last editions of the Multifarm track of OAEI evaluation campaign, we should note that the Multifarm track includes the Arabic dataset [5, 6] since 2015. Most of systems which participated at OAEI use a direct translation-based matching approach.

Table 1 summarizes the results of the systems achieving the best results in the Multifarm track in previous edition. Note that some of these systems participated in different editions and they obtained low results due to problems such as parsing or accessing to translator server. These results also includes the changes that have been made on Multifarm track. The purpose of these selection is to observe the best results obtained on Multifarm track.

The AUTOMSv2 system [14] uses a free Java API named WebTranslator[2] in order to solve the multi-language problem by translating label and properties in English language. The GOMMA system [15] uses a free translation API named "mymemory"[3] to automatically translate non-English terms. The WeSeE-Match system [16] translates the fragments, labels, and comments in English as a pivot language using the Bing[4] Search APIs translation capabilities. The WikiMatch system [17] employs the

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

[2] `http://webtranslator.sourceforge.net/`

[3] `http://mymemory.translated.net/`

[4] `https://www.microsoft.com/en-us/translator/translatorapi.aspx`

Table 1: Top systems in the multifarm track

OAEI	Top Systems	Multifarm Track	Precision	F-measure	Recall
2012	AUTOMSv2	without Arabic	0.49	0.36	0.10
2012	WeSeE	//	0.61	0.41	0.32
2012	GOMMA	//	0.29	0.31	0.36
2012	WikiMatch	//	0.34	0.27	0.23
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
2015	AML		0.53	0.51	0.50
2015	LogMap		0.75	0.41	0.29
2015	XMap		0.23	0.25	0.28
2015	CLONA		0.46	0.39	0.35
2016	CroLOM		0.55	0.36	0.28
2017	KEPLER		0.43	0.31	0.25
2017	Wikiv3		0.30	0.25	0.21

Google Translation API[5] for addressing multi-lingual ontologies. The CLONA system [18] translates the entities described in different natural languages into English as a pivot language using Bing translator. Then it uses Lucene search engine and WordNet to determine alignment candidates. The XMap system [7] 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 [8] 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 that participated in the OAEI 2014 campaign used a multilingual module based on Google translate; however the new version of the LogMap system uses both Microsoft and Google translator APIs [11]. The YAM++ system [9] uses a multilingual translator based on Microsoft Bing to translate the annotations to English. The KEPLER and Wikiv3 systems participated for the first time at OAEI2017....

We have also observed that, at OAEI2017 the best results are still those obtained by AML system in 2015, achieving an F-measure equals to 0.51. This is surprising, in spite of many research works that have been established in the field of multilingual ontology matching.

[5] `http://code.google.com/apis/language/translate/overview.html`

3 CroLOM: Cross-Lingual Ontology Matching System

We summarize the process of our approach to provide a general idea of the proposed solution. It consists in the following successive phases:

3.1 Extraction and Normalization

CroLOM extracts first the entities of the input ontologies. Then, it employs NLP to normalize the entities described in different natural languages. Unlike existing approaches, we have applied lemmatization, stemming and stopword elimination for each natural language separately before translation step. First, for each language considered by multifarm, we have established the stop words of each language in order to eliminate them from entities labels. Second, we have developed morphological algorithms to obtain lemmatization of the entities words.

This step is important [6], since one of matchers used is (1) based on string comparison algorithm to compute similarity and (2) the categories of the words are stoked in lemma form.

3.2 Translation

Once the entities are normalized, CroLOM uses the Yandex translator in order to translate the entities described in different natural languages in English as a pivot language. After translation, CroLOM employs for the second time the normalization step in order to eliminate the stop words of the English language from entities labels.

We have mentioned before that the translation path and the used translator play important role to resolve the multilingualism heterogeneity problem. Our choice for the Yandex translator is justified by the fact that it is ranked as the 4th largest search engine in the world and it has not previously used to align multilingual ontologies. However, we have chosen English as a pivot language because there a lot dictionaries that are available in English language. These dictionaries could be exploited in order to improve our system in the future. In addition, to compute the similarity between entities, we have used dictionaries (word categories and WordNet) in English. Due to automatic translation, we have observed that some stop words can be appeared in translated entities. For this purpose, we have employed the normalization for the second time.

3.3 Similarity Computation

Once the translation and standardization are carried out, CroLOM applies first, a case conversion by converting all entities words in lower case then it passes to the similarity computation step. Unlike existing systems, which use well known matchers, we have developed a matcher which calculates the similarity between entities based on the categories of the Words, string-based algorithm and synonyms using Wordnet[7].

[6] This step allows to obtain good results such as the results of our previous work [19] (STRIM system) in instance matching.

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

The matcher developed establishes a Cartesian product between the two entities words, then it returns the maximum similarity value using Levenshtein distance, similarity based on WordNet and similarity based on the categories of the words. The similarity based on the categories of the words has been adapted with some modification from the project "Calculate Semantic Similarity" [8]. The project has been developed to match sentences, however we have modified the code in order to compute similarity between words.

3.4 Alignment Identification

Finally, CroLOM 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.

4 Experimental Study

The results obtained by running our CroLOM system on multifarm track of OAEI2017 are the same as in OAEI2016 since we partipated with the same version. The results are available at the following website: `http://oaei.ontologymatching.org/2017/results/multifarm/index.html`.

5 Conclusion

In this paper, we have presented our CroLOM system, a cross-lingual ontology matching system. CroLOM unlike existing approaches, applies first NLP on each natural language before translation. Then, it uses the Yandex translator in order to translate all entities in English as pivot language. Finally, CroLOM computes the similarity between translated entities based on the category of the words and WordNet, hybridizing statistic and semantic similarity.

As future challenges, we aim to (1) improving the quality results of our system and especially the execution time, (2) conduct a survey study that addresses all the issues mentioned above, (3) taking into account the indirect approaches.

References

1. 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.
2. A. Khiat and M. Benaissa, "Boosting Reasoning-Based Approach by Structural Metrics for Ontology Alignment". The Journal of Information Processing Systems (JIPS), 2015.

[8] `https://sourceforge.net/projects/semantics/`

3. S Zhang and O. Bodenreider, "Alignment of Multiple Ontologies of Anatomy: Deriving Indirect Mappings from Direct Mappings to a Reference", AMIA 2005 Symposium Proceedings, 2005.
4. J. J. Jung, A. Hakansson, and R. H. . "Indirect Alignment between Multilingual Ontologies: A Case study of Korean and Swedish Ontologies," in Proceedings of the Third KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications,2009.
5. 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.
6. 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.
7. W. Djeddi, M. T.Khadir and S. Ben-Yahia, "XMap++ results for OAEI 2015". In Proceedings of the 10th International Workshop on Ontology Matching ISWC 2015, USA, 2015.
8. D. Faria, C. Martins, A. Nanavaty, D. Oliveira, B. Sowkarthiga, A. Taheri, C. Pesquita, F. Couto and I. Cruz , "AML results for OAEI 2015". In Proceedings of the 10th Workshop on Ontology Matching ISWC 2015, USA, 2015.
9. 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.
10. A. Khiat and M. Benaissa, "AOT / AOTL results for OAEI 2014". In Proceedings of the 9th International Workshop on Ontology Matching ISWC 2014, pp. 113-119, Italy, 2014.
11. E. Jiménez-Ruiz, BC. Grau, A. Solimando, V. Cross, "LogMap family results for OAEI 2015". In Proceedings of the 10th Workshop on Ontology Matching ISWC 2015, USA, 2015.
12. 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.
13. 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. 62-68, 2012.
14. K. Kotis, A. Katasonov and J. Leino, "AUTOMSv2 results for OAEI 2012", In Proceedings of the 7th Workshop on Ontology Matching ISWC 2012, USA, 2012.
15. A. Gro, M. Hartung, T. Kirsten and E. Rahm, "GOMMA results for OAEI 2012", In Proceedings of the 7th Workshop on Ontology Matching ISWC 2012, USA, 2012.
16. H. Paulheim, "WeSeE-Match results for OEAI 2012", In Proceedings of the 7th Workshop on Ontology Matching ISWC 2012, USA, 2012.
17. S. Hertling and H. Paulheim, "WikiMatch results for OEAI 2012", In Proceedings of the 7th Workshop on Ontology Matching ISWC 2012, pp., USA, 2012.
18. M. El-Abdi, H. Souid, M. Kachroudi and S. Ben-Yahia, "CLONA results for OAEI 2015", In Proceedings of the 10th Workshop on Ontology Matching ISWC 2015, USA, 2015.
19. A. Khiat, M. Benaissa and M. A. Belfdhal, "STRIM results for OAEI 2015 instance matching evaluation". In Proceedings of the 10th International Workshop on Ontology Matching co-located with the 14th International Semantic Web Conference (ISWC 2015), USA, 2015.
20. 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.