CroMatcher - Results for OAEI 2015

Marko Gulić [a], Boris Vrdoljak [b], Marko Banek [b,c,1]

[a] Faculty of Maritime Studies, Rijeka, Croatia
marko.gulic@pfri.hr

[b] Faculty of Electrical Engineering and Computing, Zagreb, Croatia
boris.vrdoljak@fer.hr

[c] Ericsson Nikola Tesla d.d., Krapinska 45, HR-10000 Zagreb, Croatia

Abstract. CroMatcher is an ontology matching system based on parallel composition of basic ontology matchers. There are two fundamental parts of the system: first, automated weighted aggregation of correspondences produced by different basic matchers in the parallel composition; second, an iterative final alignment method. This is the second time CroMatcher has been involved in the OAEI campaign. Basic improvement with respect to the previous version has been implemented in order to speed up the system.

1 Presentation of the system

CroMatcher is an automatic ontology matching system for discovering correspondences between entities of two different ontologies. This is the second version of the system. The first version [1] was presented in the OAEI campaign held in 2013. In this second version, the system architecture remained unchanged but the system implementation was modified as well as the implementation of several basic matchers in order to speed up the system. Our goal was to prepare the system for the following test sets: Benchmark, Anatomy, Conference and Large Biomedical Ontologies. The system is fully prepared for the Benchmark, Anatomy, and Conference. It is partly prepared for the Large Biomedical Ontologies (only for the 10% fragments of ontologies). We are currently working to speed up our system even more and we expect to present it in the next OAEI campaign.

1.1 State, purpose, general statement

As stated before, the architecture of the new version of the system remained unchanged according to the first version [1] from 2013. To recapitulate, CroMatcher contains several terminological and structural matchers connected through sequential-parallel

[1] Presently at Ericsson Nikola Tesla, the research was done while working at the University of Zagreb

composition. First, the terminological basic matchers are executed. These matchers are connected through a parallel composition. After the execution of terminological matchers, the weighted aggregation is performed in order to determine the aggregated correspondence results of these matchers. These aggregated results are used in the execution of the structural matchers as initial values of entity correspondences. Structural matchers are also executed independently of each other in another parallel composition. Again, weighted aggregation is performed in order to determine the aggregated correspondence results of the structural matchers. Before the final alignment, the aggregated correspondence results of the terminological matchers and the aggregated correspondences' results of the structural matchers need to be aggregated using weighted aggregation. Eventually, the method of the final alignment is executed. This method iteratively takes the best correspondences between two entities into the final alignment.

1.2 Specific techniques used

In this section, only the modified components will be described in detail. The rest of the main components are described in the first version of the system [1]. We modified some terminological and structural matchers in order to speed up the matching process. These matchers are modified for the test sets Anatomy and Large Biomedical Ontologies because the ontologies in these test sets contain a large number of entities. Our matcher first counts the number of entities. If the ontologies contain more than 1000 entities than the modified versions of some matchers are activated instead of the original versions of matchers. Furthermore, we modified one terminological basic matcher in order to read entity information from components *oboInOwl#hasRelatedSynonym* and *oboInOwl#hasDefinition*. These components are implemented within ontologies of the Anatomy test set and contain considerable information about entities. The modified basic matchers are the following:

1. **Terminological matchers:**

- Matcher that compares ID and annotation text of two entities (classes or properties) with the n-gram matcher [2] is extended in a way that also compares the text obtained from components *oboInOw#hasRelatedSynonym* and *oboInOwl#hasDefinition*. As stated before, these components are implemented within ontologies in the Anatomy test set. Our system first checks whether these components are implemented. If these components are not implemented within compared ontologies, the matcher compares only the ID and annotations like before.
- Matcher that compares textual profiles of two entities with TF/IDF [3] and cosine similarity [4] is modified for the ontologies that contain more than 1000 entities in order to speed up the matching process. A textual profile is a large text that describes an entity (text obtained from annotations of compared entity and its all sub entities) therefore the matching was very slow because the TF/IDF method need to load the text of all entities before starting comparing two entities. When a target ontology contains more than 1000 entities, a modified implemented matcher is activated. This matcher compares textual profiles of two entities with the string metric described in [5]. This metric calculates similarity based on

adjacent character pairs that are contained in both strings. This string metric is much faster than the TF/IDF method but the matching results are a bit worse than the results obtained with TF/IDF method. It is acceptable because the system performs the matching process faster enough to match ontologies with many entities.

- Matcher that compares individuals of two entities by applying TF/IDF and cosine similarity is modified for the ontologies that contain more than 1000 entities. If the ontology contain more than 1000 entities, a modified implemented matcher with string metric described in [5] is activated like in the previous basic matcher.
- Matcher that compares extra individuals of two entities with TF/IDF and cosine similarity is modified like two previous matchers in order to speed up the matching process.

2. **Structural matchers:**

- All structural matchers described in the first version of our system [1] are executed iteratively. In order to speed up the matching process, we also made modification when comparing ontologies that contain more than 1000 entities. All structural matchers are executed just once (instead of being executed iteratively many times) when comparing the ontologies with more than 1000 entities. This speeds up the matching process but decreases the quality of matching process when comparing large ontologies. In the next version of the system, our major concern will be to solve the problem of slow iterative execution of structural matchers.

2 Results

In this section, the evaluation results of CroMatcher matching system executed on the SEALS platform are presented.

2.1 Benchmark

In OAEI 2015, Benchmark includes two test sets: Biblio and Energy. In Table 1 the results obtained by running the CroMatcher ontology system can be seen.

Table 1. CroMatcher results for Benchmark test set

Test set	Recall	Precision	F-Measure	Time (s)
Energy	0.21	0.96	0.67	-
Biblio	0.82	0.94	0.88	485

The result for Biblio test set is equal to the result obtained at the OAEI 2013 campaign because the actual system is equal to the previous version of our system when the system matches ontologies that have less than 1000 entities. The execution time for Biblio test set was reduced by 50%, which is the result of the optimization of the program code. Our system achieves the best result in this test set together with the Lily system (F-measure 0.88). The Energy test set is new Benchmark test set. Our system achieves the third best result for this test set. Given the overall results of these two test sets, our system achieves the best result for the Benchmark test set. Most of the ontologies in Benchmark test set are implemented without entity annotations (label and comment) therefore it can be concluded that our system uses well the information from other ontology components in order to find alignment between two ontologies.

2.2 Anatomy

In OAEI 2015, the Anatomy test set consist of two large ontologies (mouse.owl and human.owl) that have to be matched. In Table 2 the results obtained by running the CroMatcher ontology system can be seen.

Table 2. CroMatcher results for Anatomy test set

Test set	Recall	Precision	F-Measure	Time (s)
Anatomy	0.814	0.914	0.861	569

Our system achieves the sixth best result for this test set. The result of our system (F-measure 0.861) is very close to the results of the better systems in this test set except the result of the system AML which is the only system with F-measure greater than 0.9 (0.944). The result for Anatomy test set is a bit lower than we expected. It is lower because the system activates modified basic matchers for the ontologies with more than 1000 entities and these matchers (especially non-iterative structure matchers) are not as good as the original basic matchers but they speed up the system very much. In OAEI 2013, our system did not finish to match ontologies in the Anatomy test set even after 5 hours which was the time limit for the OAEI 2013 campaign. Therefore, a little bit lower result is, in our opinion excusable in exchange for the speed of execution. However, a remaining challenge for future work is to speed up the execution of the iterative structural matcher in order to improve the matching results for Anatomy test set. Also, we have to improve the usage of the information obtained by components *oboInOwl#hasRelatedSynonym* and *oboInOwl#hasDefinition* which are not the standard component of the OWL ontology but are the standard implemented components in mouse.owl and human.owl ontologies.

2.3. Conference

In OAEI 2015, Conference test set consist of 16 small ontologies that have to be matched to each other. In Table 3 the results obtained by running the CroMatcher ontology system can be seen.

Table 3. CroMatcher results for Conference test set

Test set	Recall	Precision	F-Measure	Time (s)
Conference	0.50	0.59	0.54	183

The result for Conference test set classifies our system among the worst ontology systems for this test set. These ontologies mutually have approximate about ten exact correspondences therefore the best matching systems found about two correspondences more than our system which is not the big difference but considering the results of the Benchmark test set, we expected to have better result. Considering the implementation of these ontologies, it can be seen that all entities have the meaningful ID or label which is not the case for Benchmark test set. Therefore, in the Benchmark test set the threshold of the final alignment has low value but in Conference test set where all entities have meaningful names, we believe that the threshold needs to be higher. This is obviously one more challenge for the next version of our system.

2.4. Large Biomedical Ontologies, Multifarm, Interactive, Ontology Alignment for Query Answering and Instance matching

The system had problems with Large Biomedical Ontologies therefore we have to speed it up more before the next evaluation. For other test sets (Multifarm, Interactive, Ontology Alignment for Query Answering and Instance matching) the matching process itself needs to be modified and we did not prepare the system for these test sets.

3 General comments

We are very pleased for the opportunity to evaluate our ontology matching system on the SEALS platform and thus compare our system with other existing systems. There are many different test cases and we think that these test cases will help us make additional improvements of our system in the future.

3.1 Comments on the results

Our system shows great results in Benchmark test set again. We can be satisfied with the result of Anatomy test set but we will try to improve the system for these test sets. Moreover we will make our system capable of processing the sets for which we did not prepared it in this campaign.

3.2 Discussions on the way to improve the proposed system

We applied faster measure than TF/IDF to compare different documents of entities. We will try to solve the problem with the slow iterative structural matcher. Also, we will have to store the data about the entities in a separate file instead of java objects in order to reduce the usage of memory in the system.

4 Conclusion

The second version of the CroMatcher ontology matching system and its results were presented in this paper. The evaluation results show that CroMatcher achieved considerable results for Benchmark and Anatomy test sets. The matching process is executing much faster than the matching process in the first version of the system but there is still room for improvement considering speed of the process. Also, the system needs to be modified for the special test sets in the OAEI campaign like Instance matching or Multifarm. We will try to solve these problems and prepare the system to be competitive in all OAEI test sets next year.

References

1. Gulić, M., Vrdoljak, B.: CroMatcher - results for OAEI 2013, Proceedings of the 8th International Workshop on Ontology Matching, pp. 117–122, Sydney, Australia, 2013.
2. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, 2007.
3. Salton, G., McGill, M.H.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
4. Baeza-Yates, R., Ribeiro-Neto B.: Modern Information Retrieval. Addison-Wesley, Boston (1999)
5. Strike a match, http://www.catalysoft.com/articles/strikeamatch.html, accessed: 06.10.2015.