CroMatcher - Results for OAEI 2016

Marko Gulić[1], Boris Vrdoljak[2], Marko Banek[3]

[1] Faculty of Maritime Studies, Rijeka, Croatia
marko.gulic@pfri.hr
[2] Faculty of Electrical Engineering and Computing, Zagreb, Croatia
boris.vrdoljak@fer.hr
[3] Ericsson Nikola Tesla d.d., Zagreb, Croatia
marko.banek@gmail.com

Abstract. Ontology matching plays an important role in the integration of heterogeneous data sources that are described by ontologies. In order to find correspondences between entities of different ontologies, a matching system has to be built. CroMatcher is an ontology matching system that consists of several string and structural basic matchers. As individual basic matcher computes similarity between entities using information obtained from one or more components of the entire ontology, all individual matching results need to be aggregated in order to achieve the better final matching results of compared ontologies. The CroMatcher system uses weighted aggregation method that automatically determines the weighting factors of each basic matchers considering quality of its matching result. Also, the system uses iterative final alignment method that selects appropriate correspondences between entities of compared ontologies from the aggregated matching results. This is the third time CroMatcher has been involved in the OAEI campaign. The system is upgraded by introducing two new basic matchers that improved the matching results at this OAEI campaign. CroMatcher achieved excellent matching results for the three ontology matching tracks in which it participated.

1. Presentation of the system

1.1. State, purpose, general statement

Ontology matching is the process of finding semantic relationships or correspondences between entities of different ontologies [1]. A matching system has to be built in order to determine correspondences between entities. CroMatcher is an ontology matching system in which the matching process is carried out automatically. It supports the matching between ontologies expressed in Web Ontology Language (OWL) [2] that is recommended by W3C (World Wide Web Consortium) [3] as an international standard for ontology representation. There are several string and structural basic matcher in CroMatcher system. Each basic matcher determines similarity between entities using information obtained from one or more components of the compared ontologies, therefore matching results obtained by all basic matchers need to be aggregated in order to achieve the better final matching results. The string basic matchers, as well as the structural basic matchers, are related by parallel composition of basic matchers. First, the string basic matchers are executed. The results obtained by string basic matchers are automatically aggregated using our weighted aggregation method. These aggregated results are then used in the execution of the structural matchers as initial values of correspondences between entities. Again, the results obtained by structural basic matchers are aggregated using the weighted aggregation. Before the final alignment, the aggregated results of the string matchers and the aggregated results of the structural matchers are aggregated using the weighted aggregation. Eventually, the iterative final alignment method is executed in order to select appropriate correspondences between entities of compared ontologies from the aggregated matching results. The CroMatcher system that participated at OAEI 2016 is the third version of the system. Unlike the first two versions of the system [4, 5, 6] that have the identical architecture of matching process, a two new basic matchers are implemented into the newest version of the system. These matchers improved the matching results for the three ontology matching tracks in which CroMatcher participated in the OAEI campaign. CroMatcher is fully prepared for the *Benchmark* [7], *Anatomy* [8] and *Conference* [9] ontology tracks and produces excellent results for these tracks.

1.2. Specific techniques used

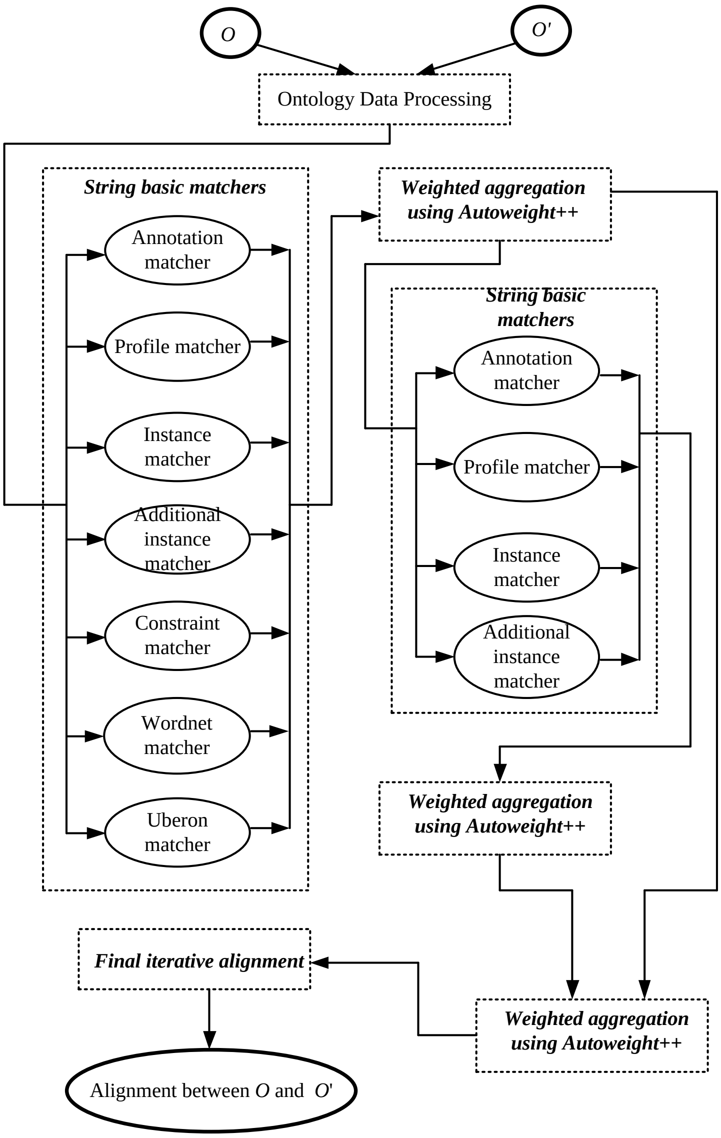
In this section, the architecture of CroMatcher system as well as the main components will be briefly presented. As already mentioned, this version of CroMatcher (OAEI campaign 2016) has two more string basic matchers implemented than last version presented in [6]. Like last year, some basic matchers are modified to speed up the matching process for *Anatomy* ontology matching track that contains a large number of entities. The system activates the lite version of these basic matchers if the compared ontologies contain more than thousand entities. The workflow and the main components of the system can be seen in the Figure 1. The CroMatcher consists of the following components:

1. **Ontology data processing** - Initial step of an ontology matching process is the extraction of information about entities within compared ontologies. After the extraction of data, the matching process starts to determine correspondences between entities of compared ontologies.

2. **String basic matchers** – determine correspondences between entities considering the character arrays (strings) that describe compared entities.

- *Annotation matcher* – determines the correspondence between entities by comparing the strings obtained from entities' IDs and annotations using n-gram similarity [1].

- *Profile matcher* - determines the correspondence between entities by comparing the textual profiles of two entities. The methods TF/IDF [10] and cosine similarity [11] are used to calculate similarity between these textual profiles.



O
O'
Ontology Data Processing
String basic matchers
Annotation matcher
Profile matcher
Instance matcher
Additional instance matcher
Constraint matcher
Wordnet matcher
Uberon matcher
Weighted aggregation using Autoweight++
String basic matchers
Annotation matcher
Profile matcher
Instance matcher
Additional instance matcher
Weighted aggregation using Autoweight++
Weighted aggregation using Autoweight++
Final iterative alignment
Alignment between O and O'

Figure 1. Workflow and the main components of CroMatcher

The textual profile is a large text that describes an entity. A content of textual profile is precisely defined in [6]. Considering the size of textual profile, the matching process is slow because the TF/IDF method has to retrieve the text of all entities before starting comparing two entities. When a target ontology contains more than 1000 entities, a modified *Profile matcher* is activated. This matcher determines correspondences using the fast string metric described in [12]. The results of this modified *Profile matcher* are a bit worse than results of the *Profile matcher* that uses TF/IDF method but it is acceptable considering the faster matching process.

- *Instance matcher* – determines the correspondence between instances of compared entities by using the methods TF/IDF and cosine similarity.
- *Additional instance matcher* - determines the correspondence between additional instances of compared entities by using the methods TF/IDF and cosine similarity. Additional instances contain not only the instances of compared entities but also the instances of entities that are related to the compared entities.
- *Constraint matcher* – determines the correspondence between entities by comparing various features of compared entities (number of object and data properties, cardinality constraints…).
- *WordNet matcher* – a newly implemented matcher. It determines the correspondence between entities by comparing the strings obtained from entities' IDs and annotations using WordNet [13]. WordNet is a large lexical database of English. The WordNet matcher can find similarities between two tokens of compared strings considering the relations (synonyms, hypernyms etc.) defined between these tokens within WordNet. The deficiency of the previous systems was its inability to recognize these language relations.
- *Uberon matcher* – a newly implemented matcher. It determines the correspondence between entities by using the mediator ontology Uberon (Uber Anatomy Ontology) [14]. This matcher is used for the Anatomy matching track. Uberon is an integrated cross-species ontology covering anatomical structures in animals. Hence, Uberon contains a lot of information about the anatomy, therefore it is very helpful when matching ontologies of the Anatomy track.

3. **Structural basic matchers** – determine correspondences between entities by comparing their relations with other entities. All these matchers are executed iteratively. Like in the previous OAEI campaign, in order to speed up the matching process, we made modified structural matchers when comparing ontologies that contain more than 1000 entities. When ontologies contain more than 1000 entities, all structural matchers are executed just once. Modified matchers decreases the quality of matching process but speed up the process.

   - *SuperEntity matcher* – determines the correspondence between entities by comparing the mutual correspondences between their parent entities.
   - *SubEntity matcher* – determines the correspondence between entities by comparing the mutual correspondences between their children entities.
   - *Domain matcher* – this matcher has two modes, one for calculating similarity between class entities and the other one for property entities. First version determines correspondences between classes by comparing all the properties that have the compared classes as their domains. Second version determines correspondences between properties by comparing the classes defined as the domain of the considered properties.

- *Range matcher* – this matcher determines correspondences only between two property entities by comparing the classes defined as the range of the considered properties.

The procedure of executing these structural matchers is described in [6] in detail.

4. **Weighted aggregation using Autoweight++ method** – As stated before, CroMatcher system executes the weighted aggregation three times during the matching process. In this system, we have introduced the Weighted aggregation that uses a new method for automatically determining the weighting factors of basic matchers. This new method determines the weighting factors of basic matchers according to the importance of the highest correspondences found within the matching results of each basic matcher. A correspondence between two entities e_i and e_j*'* is the highest correspondence if and only if it has higher value than any other correspondence of either e_i or e_j*'* with some other entity. The importance of each highest correspondence found within the matching results of a particular basic matcher is calculated comparing the complete results of this basic matcher, without taking into consideration the matching results of other basic matchers, which is the case in Autoweight++ method [6] that is used in our previous version of the system (CroMatcher 2015).
5. **Final alignment** – The final alignment method iteratively selects relevant correspondences between entities of compared ontologies. This method is presented in detail in [6].

2. Results

2.1. Benchmarks

In OAEI 2016 campaign, the *Benchmark* ontology track includes a well-known *biblio* test case. In Table 1. the results for biblio test case achieved in OAEI campaigns 2015 and 2016 by running the CroMatcher ontology system are presented.

Table 1. The matching results of CroMatcher system for Benchmark biblio test set

OAEI	**Recall**	**Precision**	**F-Measure**
2015	0.82	0.94	0.88
2016	0.83	0.96	0.89

As CroMatcher system already has achieved very good results, the improvement of the new version of the system is small, but significant. Our system achieved the best results in the *Benchmark* ontology track together with the Lily system. The introduction of the new basic matcher based on WordNet and the modified Weighted aggregation method has led to better matching results.

2.2. *Anatomy*

The Anatomy ontology track consists of two large ontologies (*mouse.owl* and *human.owl*) that have to be matched. These ontologies represent a formal description of human and mouse anatomies. In Table 2. the results for *Anatomy* ontology track achieved in OAEI campaigns 2015 and 2016 by running the CroMatcher ontology system are presented.

Table 2. The matching results of CroMatcher system for Anatomy track

OAEI	**Recall**	**Precision**	**F-Measure**	**Time (s)**
2015	0.814	0.914	0.861	569
2016	0.902	0.949	0.925	573

CroMatcher significantly improved the matching results for *Anatomy* ontology track considering the previous results of this system. The results are improved due to introducing the *Uberon* string matcher. As stated before, Uberon is an integrated cross-species ontology covering anatomical structures in animals, therefore it is very useful when determining correspondences between ontologies of the *Anatomy* track. CroMatcher achieved the second best results in the *Anatomy* track. Only the AML system has better matching results. Furthermore, only CroMatcher and AML have the F-measure higher than 0.9. However, a remaining challenge for future work is to speed up the execution of the complete system. The focus will be on the execution performance of the iterative structural matchers.

2.3. *Conference*

Conference ontology track contains 16 similar ontologies that all describe organization of a conference. The systems are evaluated according to three different modes of evaluation of which the first mode (crisp reference alignments) is the most comprehensive one. Furthermore, there exist three variants of crisp reference alignments: ra1 (the original reference alignment), ra2 (the entailed reference alignment generated as a transitive closure computed on the ra1) and ra3 (the violation free version of ra2). Each of these three variants consists of three different tests according to three different alignments between 16 conference ontologies: M1 (contains classes only), M2 (contains properties only) and M3 (contains classes and properties together). Hence, the evaluation mode crisp reference alignments produces nine different evaluation tests for matching systems: ra1-M1, ra1-M2… ra3-M3. In this section, we will present the results of these nine different evaluation tests according to standard F-measure (the harmonic mean of precision and recall). CroMatcher system produces the best results for three tests (ra1-M1, ra2-M1 and ra3-M1). For two tests (ra2-M3 and ra3-M3), our system also produces the best results alongside the AML system. Furthermore, for remained four tests (ra1-M2, ra1-M3, ra1-M2 and ra3-M2), our system produces the second best result behind the AML system. Considering the overall results of the previous and the current version of CroMatcher (Table 3.), it can be seen that we made a great improvement in matching ontologies of *Conference* track.

Table 3. The matching results of CroMatcher system for Conference track

OAEI	Recall	Precision	F-Measure
2015	0.46	0.56	0.51
2016	0.64	0.77	0.70

2.4. *Other ontology tracks*

This year, we have not participated in other ontology tracks because we did not prepare our system for these tracks. Next year, we will try to improve our system to be able to obtain the considerable matching results for more ontology tracks than this year.

3. General comments

OAEI campaign provides not only the evaluation of our system but also the comparison with other state-of-the-art system. We consider that OAEI evaluation of the ontology matching systems is the most authoritative criterion for comparing various matching system because the complete evaluation is performed publicly by the OAEI organizers. There are also many different ontology tracks and we think that these tracks can help anybody to make additional improvements of matching system.

3.1. Comments on the results

CroMatcher achieved great matching results in the ontology tracks (Benchmarks, Anatomy, Conference) for which it was prepared. Considering the results of each individual track, our system achieved the best or the second best matching results.

3.2 Discussions on the way to improve the proposed system

We will try to solve the problem with the slow iterative structural matcher in order to improve the matching process when comparing large ontologies. 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. Furthermore, we will try to prepare the system for all OAEI ontology tracks.

4. Conclusion

The third version of the CroMatcher ontology matching system and its results in the OAEI campaign were presented in this paper. As in the previous versions of the system, CroMatcher consists of several string and structural basic matchers. The Autoweight++ method is used to aggregate the results obtained by these matchers. At the end of the matching process, the iterative final alignment method is executed. In this version of the system, two new string matchers are introduced: *WordNet* matcher and *Uberon* matcher. *WordNet* matcher can find similarities between entities considering the language relations like synonyms, hypernyms etc. Uberon is an integrated cross-species ontology covering anatomical structures in animals. Considering the *Anatomy* track, Uberon is very useful when finding correspondences between ontologies of this track. The evaluation results show that CroMatcher achieved great results for *Benchmark*, *Anatomy* and *Conference* tracks for which it was prepared. According to the results of these three tracks, CroMatcher achieved better matching results than last year. Furthermore, there is still room for improvement considering the speed of the matching process. Also, we will try to prepare the system for all ontology tracks in the OAEI campaign next year.

References

1. J. Euzenat, P. Shvaiko, Ontology matching, 2nd Edition, Springer-Verlag, Heidelberg (DE), 2013.
2. G. Antoniou, F. van Harmelen, A Semantic Web Primer, MIT Press, 2004.
3. World wide web consortium, http://www.w3.org/, accessed: 25-09-2016.
4. M. Gulić, B. Vrdoljak, CroMatcher - results for OAEI 2013, in: P. Shvaiko, J. Euzenat, K. Srinivas, M. Mao, E. Jiménez-Ruiz (Eds.), Proc. of the 8th Int. Workshop on Ontology Matching co-located with the 12th Int. Semantic Web Conf. (ISWC 2013), Sydney, Australia, October 21, 2013, Vol. 1111 of CEUR Workshop Proceedings, CEUR-WS.org, pp. 117–122.
5. M. Gulić, B. Vrdoljak, M. Banek, CroMatcher results for OAEI 2015, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, O. Hassanzadeh (Eds.), Proc. of the 10th Int. Workshop on Ontology Matching collocated with the 14th Int. Semantic Web Conf. (ISWC 2015), Bethlehem, PA, USA, October 12, 2015, Vol. 1545 of CEUR Workshop Proceedings, CEUR-WS.org, 2016, pp. 130–135.
6. M. Gulić, B. Vrdoljak, M. Banek, CroMatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment, Journal of Web Semantics: Science, Services and Agents on the World Wide Web (2016), http://dx.doi.org/10.1016/j.websem.2016.09.001
7. J. Euzenat, M.-E. Rosoiu, C. Trojahn, Ontology matching benchmarks: generation, stability, and discriminability, Journal of Web Semantics, Vol 21 (2013) 30–48.
8. O. Bodenreider, T.F. Hayamizu, M. Ringwald., S. de Coronado, S. Zhang, Of mice and men: Aligning mouse and human anatomies, AMIA Annu Symp Proc, 2005, pp. 61-65
9. M. Cheatham, P. Hitzler, Conference v2.0: An Uncertain Version of the OAEI Conference Benchmark. International Semantic Web Conference (2), 2014, pp. 33-48.
10. G. Salton, M.H. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983
11. R. Baeza-Yates, B. Ribeiro-Neto, Modern Informationl Retrieval. Addison-Wesley, Boston, 1999
12. Strike a match, http://www.catalysoft.com/articles/strikeamatch.html, accessed 25-09-2016
13. G. A. Miller, WordNet: A Lexical Database for English. Communications of the ACM, 38(11):39–41, 1995.
14. C. J. Mungall, C. Torniai, G. V. Gkoutos, S. Lewis, and M. A. Haendel. Uberon, an Integrative Multi-species Anatomy Ontology. Genome Biology, 13(1): R5, 2012.
15. M. Gulić, I. Magdalenić, B. Vrdoljak, Automatically specifying parallel composition of matchers in ontology matching process, Communications in Computer and Information Science, vol. 240, 2011, pp. 22–33.