MaasMatch results for OAEI 2011

Frederik C. Schadd, Nico Roos

Maastricht University, The Netherlands
{frederik.schadd, roos}@maastrichtuniversity.nl

Abstract. This paper summarizes the results of the first participation of MaasMatch in the Ontology Alignment Evaluation Initiative (OAEI) of 2011. We provide a brief description of the techniques that have been applied, with the emphasis being on the application of virtual documents and information retrieval techniques in order effectively utilize linguistic ontologies. Also, we discuss the results achieved in the tracks provided under the SEALS modality: benchmark, conference and anatomy.

1 Presentation of the system

1.1 State, purpose, general statement

Sharing and reusing knowledge is an important aspect in modern information systems. Since multiple decades, researchers have been investigating methods that facilitate knowledge sharing in the corporate domain, allowing for instance the integration of external data into a company's own knowledge system. Ontologies are at the center of this research, allowing the explicit definition of a knowledge domain. With the steady development of ontology languages, such as the current OWL language [3], knowledge domains can be modeled with an increasing amount of detail.

Unfortunately, since ontologies of the same knowledge domain are commonly developed separately or for different purposes, transferring information across different sources becomes challenging as the heterogeneities between the ontologies need to be resolved. Several types of heterogeneities can emerge between two ontologies, commonly divided into syntactic, terminological, semantic and semiotic heterogeneities [1].

MaasMatch is an ontology matching tool with a focus on resolving terminological heterogeneities, such that entities with the same meaning but differing names and entities with the same name but different meanings are identified as such and matched accordingly. Given this focus, the tool has been primarily tested using the conference data set, since the ontologies of this data set are more likely to contain these heterogeneities.

1.2 Specific techniques used

In this section we will present the techniques applied in MaasMatch. The overall structure of MaasMatch is simple, being a combination of a string similarity measure and our WordNet similarity, and using the combination of the two similarity matrices to extract the final alignments. However, most of our research so far has been invested into advancing the effectiveness of WordNet similarities.

WordNet makes it possible identify concepts that have the same meaning but different names, since synonyms are grouped into sets, called synsets. However, a more challenging task is the identification of concepts have similar names, but different meanings. As an example, if an ontology contains a concept 'house', then WordNet contains 14 different meanings for this word, and hence 14 different synsets that can be described by this name. One is thus faced with the challenge of automatically identifying the synset that denotes the correct meaning of the ontology entity. To do this, we applied a combination of information retrieval techniques and the creation of virtual documents in order to determine which synset most likely denotes the correct meaning of an entity. That way, only synsets which resulted in a high document similarity with their corresponding concept are subsequently used for the calculation of the WordNet similarity.

The approach can be separated into 5 distinct steps as follows: Given two ontologies O_1 and O_2 that are to be matched, where O_1 contains the sets of entities $E_x^1 = \{e_1^1, e_2^1, ..., e_m^1\}$, where x distinguishes between the set of classes, properties or instances, and O_2 contains the sets of entities $E_y^2 = \{e_1^2, e_2^2, ..., e_n^2\}$, and where $C(e)$ denotes a collection of synsets representing entity e, the main steps of our approach, performed separately for classes, properties and instances, can be described as follows:

1. **Synset Gathering:** For every entity e in E_x^i, assemble the set $C(e)$ with synsets that might denote the meaning of entity e.
2. **Virtual Document Creation:** For every entity e in E_x^i, create a virtual document of e, and a virtual document for every synset in $C(e)$.
3. **Document Similarity:** For every entity e in E_x^i, calculate the document similarities between the virtual document denoting e and the different virtual documents originating from $C(e)$.
4. **Synset Selection:** For every collection $C(e)$, discard all synsets from $C(e)$ that resulted in a low similarity score with the virtual document of e, using some selection procedure.
5. **WordNet Similarity:** Compute the WordNet similarity for all combinations of $e^1 \in E_x^1$ and $e^2 \in E_x^2$ using the processed collections $C(e^1)$ and $C(e^2)$.

The first step of the procedure is fairly straightforward, where all corresponding synsets are collected if the complete name of an entity is present in WordNet and string processing techniques such as word stemming or finding legal sub-strings in the name are applied if the complete name is not present in WordNet. Figure 1 illustrates steps 2 - 5 of our approach for two arbitrary ontology entities e^1 and e^2:

Once the similarity matrix, meaning all pairwise similarities between the entities of both ontologies, are computed, the final alignment of the matching process can be extracted or the matrix can be combined with similarity matrices stemming from other approaches.

Virtual Documents The second step of the approach consists of the creation of virtual documents for an ontology entity and several synsets that might denote the actual meaning of the entity. When constructing the virtual document, one must collect information from the ontology, or WordNet if a virtual document of a synset is constructed, in such a way that the resulting document adequately describes the meaning of the entity. An

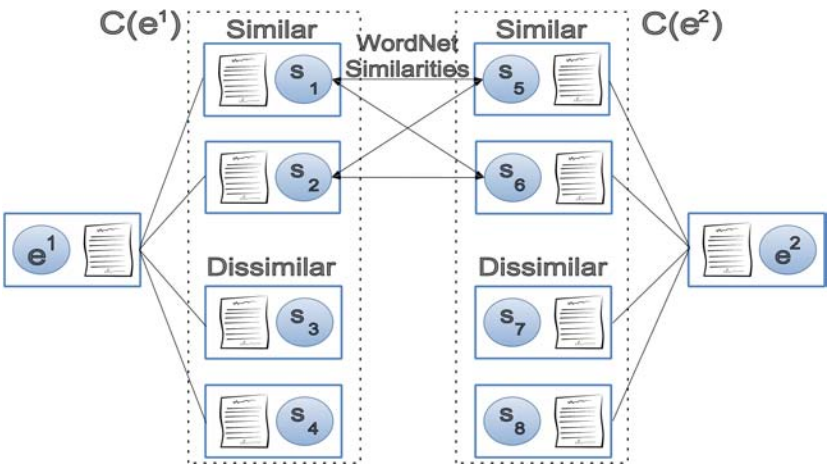

Fig. 1. Visualization of step 2-5 of the proposed approach for any entity e^1 from ontology O_1 and any entity e^2 from ontology 2.

expressive ontology such as OWL allows for the collection from various sources of information. In addition to its own name, an entity can also contain comments, which usually are written descriptions of the entity, and multiple labels. Providing context information is also beneficial. To do this, the names of the parent and child entities are also added to the document. Different details are added given the different types of entities. For virtual documents of classes, the names of all its properties are added and for properties the names of all the classes inside their range and domain are added.

Once all the information for a virtual document is collected, several post-processing techniques such as word-stemming and stop-word removal are applied, before the document is transformed into the vector-space model. Using the document vectors, the similarity between the entity document and the different synset documents is then computed using the cosine similarity.

Synset Selection Based on the document similarities between an entity and the potential synsets, some of the synsets are then discarded based on their similarity value. Several selection procedures have been tested, for instance using a cut-off value by computing the arithmetic or geometric mean of the similarities. Another tested method consisted of retaining only synsets whose document similarity had a higher value than the sum of the mean and standard deviation of the similarity value, which had the intriguing property that only few synsets remained if their similarity values were distinctly higher than the others, and more if this wasn't the case and thus it was uncertain which synset was actually appropriate. Experimentation revealed that stricter selection methods performed better than lenient methods, with the simple method of using only the synset with the highest document similarity to compute the WordNet distance resulting in the highest scoring alignments [4].

1.3 Adaptations made for the evaluation

For experiments unrelated to the actual OAEI competition, but using some of the OAEI datasets, a cut-off confidence value of 0.7 has been used for the alignments, since this is one of the standard values that has previously been used for OAEI evaluations. For the purpose of OAEI participation, however, this value was altered to 0.95 to improve the final F-Measures achieved by the system, especially for the conference data set.

1.4 Link to the system and parameters file

MaasMatch and its corresponding parameter file is available on the SEALS platform and can be downloaded at http://www.seals-project.eu/tool-services/browse-tools.

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

The computed alignments for the preliminary evaluation can be found at http://www.personeel.unimaas.nl/frederik-schadd/MaasMatchOAEI2011results.zip.

2 Results

This section presents the evaluation of MaasMatch using the benchmark, anatomy and conference data sets.

2.1 Benchmark

The results of the benchmark data set are grouped into several categories, such that benchmarks that test similar aspects of the matching process are grouped together. These can be viewed in table 1.

Ontologies	Precision	Recall	F-Measure
101-104	1.00	0.989	0.994
201-210	0.940	0.375	0.536
221-231	0.987	0.965	0.976
232-247	0.977	0.899	0.936
248-266	0.722	0.279	0.403
301-304	0.930	0.472	0.626
total	0.815	0.427	0.561

Table 1. Results of the benchmark data set.

Initially, we can see that MaasMatch produces very good results for the groups 101-104, 221-231 and 232-247. These groups have in common that they generally test alterations of aspects such as the ontology structure, instances or properties. While these

alterations can have some effects on the construction of the virtual document, as for instance the descriptions of parent classes will not be included if the hierarchy is flattened, the overall performance decreases only marginally. The groups 201-210 and 248-266, however, test alterations of the entity names and comments, resulting in a poorer performance and especially impacting the recall of the alignments. The suppression of the entity comments causes the quality of the virtual documents to drop considerably, making it harder to locate appropriate synsets in WordNet. However, most importantly, the WordNet matcher relies on applying string analysis techniques on the entity names to locate potential synsets in WordNet. If these names are altered severely, scrambled or even completely omitted, then it becomes exceedingly difficult to locate potential synsets in WordNet, which would require more sophisticated approaches than currently applied. The benchmark group 301-304, in which the default ontology is matched against several real-world ontologies, displays a similar trend regarding precision and recall, albeit less severe. Overall, one can conclude that MaasMatch achieved a fairly high precision and a moderate recall value across the entire benchmark data set.

2.2 Anatomy

The anatomy data set consists of two large real-world ontologies from the biomedical domain, with one ontology describing the anatomy of a mouse and the other being the NCI Thesaurus, which describes the human anatomy. The results of this data set can be seen in table 2.

Ontologies	Precision	Recall	F-Measure
mouse-human	0.988	0.284	0.442

Table 2. Results of the anatomy data set.

We can see that MaasMatch achieved a high precision and low recall value. The low recall value can be explained by the fact that WordNet does not contain definitions of highly technical medical terms, resulting in the system being unable to match entities which are not located in the WordNet database. Using a different linguistic ontology should alleviate this problem, or ideally the system should automatically select the most appropriate linguistic ontology for this task.

2.3 Conferences

The conference data set, which so far has been the focus during the development of MaasMatch, consists of 7 real-world ontologies that all describe the domain of organizing conferences. Here, all possible combinations of ontologies are matched and evaluated. The results of each combination can be seen in table 3.

The results of the evaluations vary. Most alignments exhibit a similar trend as in the previous two data sets, being a high precision and moderate recall. However, there are

Ontologies	Precision	Recall	F-Measure
cmt-confOf	0.800	0.250	0.380
cmt-conference	0.800	0.250	0.380
cmt-edas	0.888	0.615	0.727
cmt-ekaw	0.833	0.454	0.588
cmt-iasted	0.800	1.000	0.888
cmt-sigkdd	0.800	0.666	0.727
confOf-edas	0.538	0.368	0.437
confOf-ekaw	0.888	0.400	0.551
confOf-iasted	0.800	0.444	0.571
confOf-sigkdd	1.000	0.428	0.599
conference-confOf	0.750	0.400	0.521
conference-edas	0.857	0.352	0.499
conference-ekaw	0.727	0.320	0.444
conference-iasted	0.800	0.285	0.421
conference-sigkdd	0.875	0.466	0.608
edas-ekaw	0.666	0.260	0.374
edas-iasted	0.857	0.315	0.461
edas-sigkdd	1.000	0.466	0.636
ekaw-iasted	0.857	0.600	0.705
ekaw-sigkdd	1.000	0.636	0.777
iasted-sigkdd	0.785	0.733	0.758
total	0.825	0.462	0.592

Table 3. Results of the conference data set.

a few exceptions. A few alignments, such as cmt-iasted or iasted-sigkdd, have a higher than moderate recall. Conversely, alignments such as confOf-edas or edas-ekaw have a lower than usual precision. Overall, we can see that the aggregated performance is competitive when compared against the results of the OAEI 2010 participants [2].

3 General comments

3.1 Comments on the results

The first entrance of MaasMatch has shown a promising performance over the several data sets, most notably due to high precision values under various scenarios. With the focus being on the conference data set, the overall f-measure of MaasMatch has been established at a solid 0.592.

3.2 Discussions on the way to improve the proposed system

Several opportunities present themselves for further improvements. First, as observed during the benchmark experiments, the current approach suffers when ontologies use entity names where potentially appropriate synsets cannot be located in WordNet, due to severely altered or scrambled names. So far, processing techniques such as tokenization,

word stemming and stop word removal have been deployed to alleviate this problem, however in severe cases this problem still persists. Hence, it would be beneficial to look into further techniques that enables the matcher to locate synsets even if the entity names are scrambled.

Also, as evident in the anatomy results, if the domain to be modeled contains concepts that are not defined in WordNet, the system is unable to match it correctly. Switching WordNet with a more appropriate linguistic ontology can alleviate this problem. However, this would require the existence of such an ontology for that specific domain. Another predicament is that the system either ceases to be fully automatic, as one would need to manually define an appropriate linguistic ontology for each matching process, or would require the ability to automatically identify an appropriate linguistic ontology each time before the matching is performed.

Given the focus on resolving terminological heterogeneities, one obvious improvement would be the addition of other techniques such that other types of heterogeneities are also identified and resolved, such that the overall recall of the system is improved.

3.3 Comments on the SEALS platform

The SEALS platform is a promising project for researchers to share, test and evaluate their tools. It is fairly easy to integrate a tool into the platform and provides a neutral common ground for a comparison of systems. Unfortunately, two still absent functionalities prevent the platform of becoming the central pivot of ontology matching research. One would be the functionality for the tool developers to independently initiate the testing and evaluation of a tool using the platform, allowing a more speedy development of the tool. Secondly, the standard automatic evaluation function only computes precision and recall for the separate matching tasks. However, previous OAEI competitions also aggregated the results of different matching tasks using a form of the harmonic mean. Since the harmonic mean is not defined for values of 0, which can occur when computing the precision, recall or f-measure of an alignment, a few tweaks are necessary such that this measure can still be applied. Having the results of the separate tasks aggregated automatically by the SEALS platform would eliminate discrepancies that could arise by slightly differing implementations of such aggregation measures, so that it is ensured that the overall performances of the different tools can legitimately be compared. Overall, we are looking forward to the future developments of this platform and possibly utilizing it for further research.

4 Conclusion

This paper presented the results of the first participation of MaasMatch in the Ontology Alignment Evaluation Initiative competition. We described the techniques that are applied in the system and presented the achieved results in the benchmark, anatomy and conference data sets, with emphasis on the conference data set which has been the focus for the development of the system. We note encouraging results and discuss strengths, weaknesses and possible improvements of the system.

References

1. J . Euzenat. Towards a principled approach to semantic interoperability. In A. Gómez Pérez, M. Gruninger, H. Stuckenschmidt, and M. Uschold, editors, *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, pages 19–25, 2001.
2. J. Euzenat, A. Ferrara, C. Meilicke, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, and C. Trojahn. First results of the ontology alignment evaluation initiative 2010. In *Proceedings of ISWC Workshop on OM*, 2010.
3. D. L. McGuinness and F. van Harmelen. OWL web ontology language overview. W3C recommendation, W3C, February 2004.
4. F. Schadd and N. Roos. Improving ontology matchers utilizing linguistic ontologies: an information retrieval approach. In *Proceedings of the 23rd Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2011)*, 2011. Accepted Paper.