Matching Ontologies Through Siamese Neural Network

Xingsi Xue[1,2(✉)], Chao Jiang[1,2], and Hai Zhu[3]

[1] Intelligent Information Processing Research Center, Fujian University of Technology, Fuzhou 350118, Fujian, China
[2] School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, Fujian, China
[3] School of Network Engineering, Zhoukou Normal University, Zhoukou 466001, Henan, China

Abstract. Ontology, the kernel technique of Semantic Web (SW), formally names the domain concepts and their relationships. However, as the ontologies are created and developed by different domain experts and communities, a concept may be named in various ways, bringing about the concept heterogeneity problem. To solve the problem, in this paper, a Siamese Neural Network (SNN)-based Ontology Matching Technique (OMT) is proposed, which is able to improve the matching efficiency by using a part of Reference Alignment (RA) to decrease the training time and improve the quality of matching results by using a logic reasoning approach to remove the conflict correspondences. The experimental results demonstrate that SNN-based OMT can determine high-quality alignment which outperforms the state-of-the-art OMTs.

Keywords: Ontology matching · Siamese neural networks · OAEI

1 Introduction

Over the past decades, Semantic Web (SW) technologies have been widely utilized, which provides great convenience for people to handle and link a variety of data [1,3,18,20]. Ontology, the kernel technique of SW, formally names the domain concepts and their relationships. However, as the ontologies are created and developed by different domain experts and communities, a concept may be named in various ways, bringing about the problem of concept heterogeneity. To address this heterogeneity problem, it is vital to determine correspondences between heterogeneous concepts, called the Ontology Matching (OM) [23].

Due to the complication of OM, it is arduous and impracticable to manually establish correspondences. Hence, diverse (semi)automatic OM Techniques (OMTs) have been proposed [8,12,19,24,26]. Machine learning (ML) is widely used in various fields [4–7,14–16]. In particular, ML-based OMTs are deemed as promising methods. Doan et al. [9] first presented an ML-based OMT that

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2021
Published by Springer Nature Switzerland AG 2021. All Rights Reserved
J. Xiong et al. (Eds.): MobiMedia 2021, LNICST 394, pp. 715–724, 2021.
https://doi.org/10.1007/978-3-030-89814-4_52

similarity measures were described by a joint probability distribution of concepts concerned. Mao et al. [17] regarded the OM as a binary classification problem and used the Support Vector Machine (SVM) to solve it. khoudja et al. [13] integrated the state-of-the-art ontology matchers through the neural network to improve the results' quality. Bento et al. [2] used convolutional neural networks (CNN) to carry out the OM which shows good performance. Jiang et al. [11] proposed a Long Short-Term Memory Networks (LSTM)-based OMT to matching biomedical ontologies by using the semantic and structural information of concepts. However, the three neural network-based approaches need whole Reference Alignment (RA), which is unrealistic to obtain in the real matching scene, and three models' training is time-consuming, which could reduce the matching efficiency. In this paper, a Siamese Neural Network (SNN)-based OMT is proposed to further enhance the quality of alignments, which can predict the similarity value of two concepts by capturing the semantic feature. In particular, SNN-based OMT just utilizes a small part of RA, which is able to determine excellent alignments, and decrease the training time, and use a logic reasoning approach to remove the conflict correspondences to improve the quality of matching result.

2 Preliminary

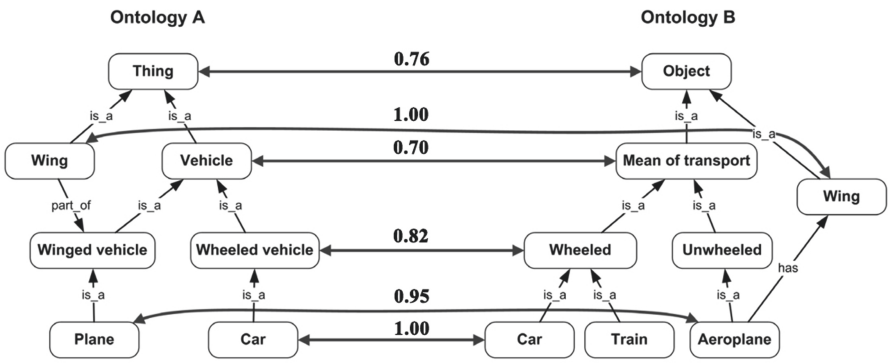

Fig. 1. Two ontologies and their alignment.

2.1 Ontology and Ontology Matching

Definition 1. An ontology is defined as a quadruple [21]

$$O = (IN, OP, DP, CL) \tag{1}$$

where IN is a series of individuals; OP is a series of object properties; DP is a series of data properties; CL is a series of classes. In addition, IN, OP, DP, and CL are the concepts. Figure 1 shows an example of two ontologies and their

alignment. Two ontologies A and B contain a number of concepts in the left and right, respectively. In addition, the elements of rounded rectangles are classes, e.g. "car", while one-way arrows are object properties, e.g. "is_a".

Definition 2. An ontology alignment is a number of correspondences, and a correspondence is a quadruple [25]

$$corres = (c', c, s, t) \tag{2}$$

where c' and c are the concept from two to-be-matched ontologies; s is the similarity in [0, 1]; and t is the type of relation between c' and c. In Fig. 1, the double-sided arrows connect two concepts constituting the correspondences. For instance, "Plane" in ontology A and "Aeroplane" in ontology B are connected building a correspondence with 0.95 similarity. Besides, all the correspondences form an alignment, and RA is a golden alignment provided by the domain experts, which is used to test the performance of OMTs.

Definition 3. The process of ontology matching is a function [22]

$$A = \phi(O_1, O_2, RA, R, P) \tag{3}$$

where A is the final alignment; O_1 and O_2 are two to-be-aligned ontologies; RA is the reference alignment; R is the used resources; P is the used parameters. In the process of ontology matching, it is vital to determine the correspondences among heterogeneous concepts.

2.2 Performance Metrics

Generally, three metrics, i.e. precision, recall, and f-measure, are utilized to test the matching results' quality [10], which are expressed as follows:

$$P = \frac{correct\_matched\_correspondences}{all\_matched\_correspondences} \tag{4}$$

$$R = \frac{correct\_matched\_correspondences}{all\_correct\_correspondences} \tag{5}$$

$$F = 2 \times \frac{P \times R}{P + R} \tag{6}$$

where P and R respectively represent the accuracy and completeness of the results. P equals 1 indicating all matched correspondences are correct, while R equals 1 meaning that all correct correspondences are matched; F is the harmonic mean of P and R to balance them.

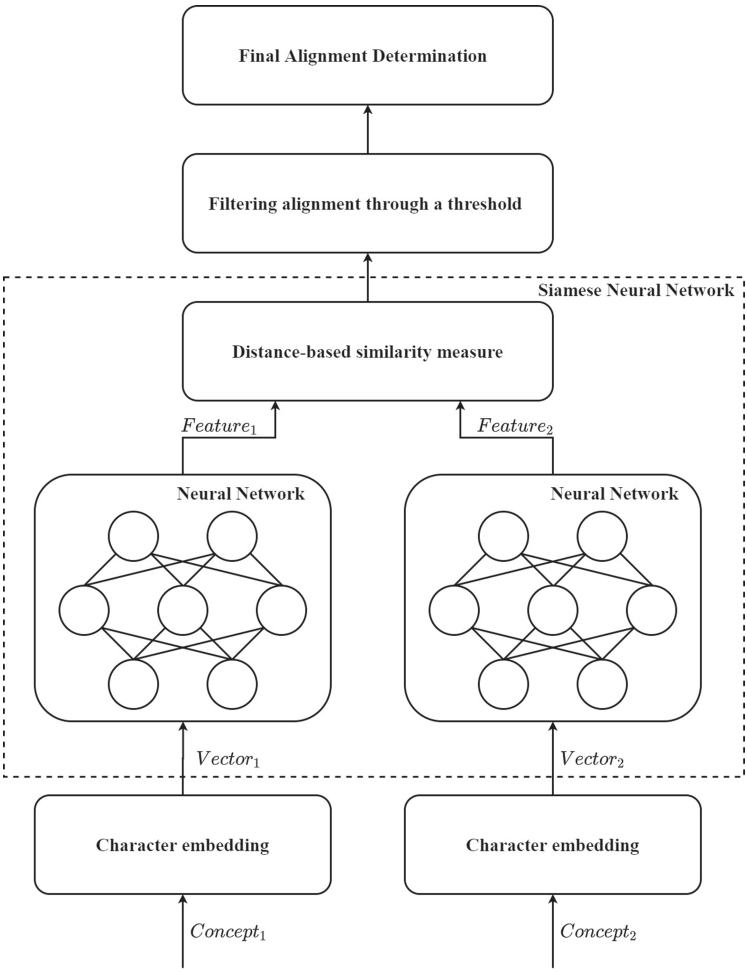

Fig. 2. The framework of the SNN-based OMT.

3 Methodology

In this study, an SNN-based OMT is proposed to solve the concept heterogeneity problem and the framework of SNN-based OMT is depicted in Fig. 2.

3.1 Data Set

The well-known Ontology Alignment Evaluation Initiative (OAEI) benchmark[1] is utilized to train the model and test the performance of our proposal. The succinct statement of benchmark testing cases is demonstrated in Table 1. To

[1] http://oaei.ontologymatching.org/2016/benchmarks/index.html.

enhance the efficiency of training process, a small part of RA is used that only a 2xx testing case is selected to train the model. In particular, given two to-be-matched ontologies O_1 and O_2, the correspondence $corres = (c', c)$ in RA is the positive sample, then the same number of negative samples are constituted by replacing c' with a concept c'' randomly selected from O_1. In addition, to ensure the quality of the final result, the trained testing case is not utilized to test.

Table 1. The succinct statement on benchmark testing cases.

ID	Succinct statement
1XX	Two same ontologies
2XX	Two ontologies with different lexical, linguistic or structural characters
3XX	The real ontologies

3.2 Siamese Neural Network

As seen in Fig. 2, the concepts' information is transformed as numeric vectors to feed the neural networks by using the character embeddings[2], whose possible character is a representation vector in 300 dimensions, and the value in each dimension is normalized in the interval [0, 1].

After, the trained SNN is utilized whose two networks are the same structure and weights. In this paper, the structure of two networks is Bi-directional Long Short Term Memory (Bi-LSTM), which is able to capture the semantic relationships and features of concept pairs. Then the distance-based similarity value of two concepts can be calculated through $Feature_1$ and $Feature_2$, which is defined as follows:

$$distance = \frac{||Feature_1 - Feature_2||}{||Feature_1|| + ||Feature_2||} \tag{7}$$

$$similarity = 1 - distance \tag{8}$$

where $||\cdot||$ represents the Euclidean norm, then the normalized distance is utilized to compute the similarity value, i.e. the smaller the distance, the greater the similarity value. In particular, the Adam optimizer and the contrastive loss are adopted and to optimize the two same networks:

$$L_{contrastive} = \sum_{i=0}^{N} \frac{y_i \times d_i + (1 - y_i) \times max\{margin - d_i, 0\}}{2N} \tag{9}$$

where N is the quantity of samples; $margin$ is a margin value; y_i is the sample type that y_i equals 1 representing the positive sample and 0 denoting the negative sample; d is the distance through the Eq. 7.

[2] https://github.com/minimaxir/char-embeddings.

3.3 The Alignment Determination

By means of the trained SNN, the $M \times N$ similarity matrix can be obtained that M and N are the concepts number of O_1 and O_2, respectively. Each element in the similarity matrix is the similarity value computed by using the Eqs. 7 and 8. In addition, the maximal value in each row or column will be chosen as the final correspondence. Furthermore, to ensure the precision of the final alignment, a threshold is adopted to filter the correspondences. After that, a logic reasoning approach is utilized that: (1) the correspondences are sorted by descending, (2) the correspondence with the greatest similarity is selected, (3) the rest correspondences are chosen one by one if it does not conflict with previous correspondence. To be specific, in Fig. 3, two correspondences (c'_1, c_1) and (c'_2, c_3) conflict that c'_1 is the subclass of c'_2 and c_3 is the subclass of c_1, then the correspondence with the lower similarity score will be discarded.

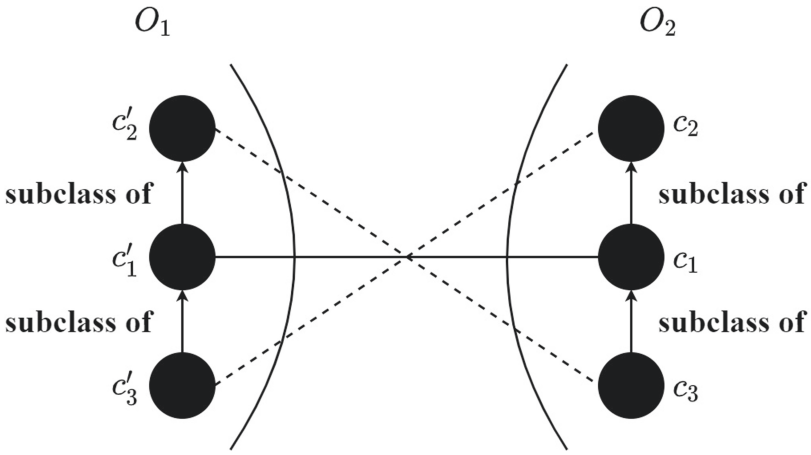

Fig. 3. An example of inconsistent correspondences.

4 Experiment

In the conducted experiment, the OAEI benchmark is used for evaluating the quality of alignments. Table 2 shows the comparison in terms of the average matching results' quality among SNN-based OMT and OAEI's OMTs, i.e. XMap, LogMap family, Pheno family, Lily, CroMatcher, and AML. As shown in Table 2, SNN-based OMT's recall, precision, f-measure, and f-measure per second are better than other competitors, which shows that SNN-based OMT outperforms the state-of-the-art OMTs.

In addition, Tables 3, 4, and 5 demonstrate the comparison in terms of R, P, and F, respectively. The bold text indicates the best OMT on the corresponding testing case and the symbols "+", "=", "-" represent the numbers that our proposal is superior, equal, and inferior to other OMTs, respectively. The compared

Table 2. Comparison among SNN-based OMT and OAEI's OMTs in terms of the average matching results' quality.

OMT	Recall	Precision	f-measure	Runtime (second)	f-measure per second
XMap	0.40	0.95	0.56	123	0.0045
PhenoMP	0.01	0.02	0.01	1833	0.0000
PhenoMM	0.01	0.03	0.01	1743	0.0000
PhenoMF	0.01	0.03	0.01	1632	0.0000
LogMapBio	0.24	0.48	0.32	54439	0.0000
LogMapLt	0.50	0.43	0.46	96	0.0048
LogMap	0.39	0.93	0.55	194	0.0028
Lily	0.83	0.97	0.89	2211	0.0004
CroMatcher	0.83	0.96	0.89	1100	0.0008
AML	0.24	1.00	0.38	120	0.0031
SNN-based OMT	0.88	0.97	0.92	108	0.0085

Table 3. Comparison among SNN-based OMT and OAEI's OMTs in terms of recall.

OMT	1XX	2XX	3XX
edna	**1.00**	0.56	0.82
aflood	**1.00**	0.74	0.81
AgrMaker	0.98	0.60	0.79
aroma	**1.00**	0.69	0.78
ASMOV	**1.00**	**0.85**	0.82
DSSim	**1.00**	0.62	0.67
GeRoMe	**1.00**	0.71	0.60
kosimap	0.99	0.57	0.50
Lily	**1.00**	0.86	0.81
MapPSO	**1.00**	0.73	0.29
RiMOM	**1.00**	0.81	0.82
SOBOM	0.97	0.46	0.55
TaxoMap	0.34	0.23	0.31
SNN-based OMT	**1.00**	0.81	**0.83**
+/=/-	4/9/0	12/0/1	13/0/0

OMT are edna, aflood, AgrMaker, aroma, ASMOV, DSSim, GeRoMe, kosimap, Lily, MapPSO, RiMOM, SOBOM, and TaxoMap. About testing cases 1XX, most of OMTs are able to obtain excellent alignments as the low heterogeneity of 1xx, that information of all concepts of ontologies have not been discarded and changed. Regarding 2XX, since the to-be-matched ontologies have a variety

Table 4. Comparison among SNN-based OMT and OAEI's OMTs in terms of precision.

OMT	1XX	2XX	3XX
edna	0.96	0.41	0.47
aflood	**1.00**	**0.98**	0.9
AgrMaker	0.98	0.98	0.92
aroma	**1.00**	**0.98**	0.85
ASMOV	**1.00**	0.96	0.81
DSSim	**1.00**	0.97	0.94
GeRoMe	**1.00**	0.92	0.68
kosimap	0.99	0.94	0.72
Lily	**1.00**	0.97	0.84
MapPSO	**1.00**	0.75	0.54
RiMOM	**1.00**	0.93	0.81
SOBOM	0.98	0.97	0.92
TaxoMap	**1.00**	0.9	0.77
SNN-based OMT	**1.00**	**0.98**	**0.94**
+/=/-	4/9/0	11/2/0	13/0/0

Table 5. Comparison among SNN-based OMT and OAEI's OMTs in terms of f-measure.

OMT	1XX	2XX	3XX
edna	0.98	0.47	0.59
aflood	**1.00**	0.84	0.85
AgrMaker	0.98	0.74	0.85
aroma	**1.00**	0.80	0.81
ASMOV	**1.00**	0.90	0.81
DSSim	**1.00**	0.75	0.78
GeRoMe	**1.00**	0.80	0.63
kosimap	0.99	0.7	0.59
Lily	**1.00**	**0.91**	0.82
MapPSO	**1.00**	0.74	0.37
RiMOM	**1.00**	0.86	0.81
SOBOM	0.97	0.62	0.68
TaxoMap	0.50	0.36	0.44
SNN-based OMT	**1.00**	0.88	**0.88**
+/=/-	5/8/0	11/0/2	13/0/0

of heterogeneity features, i.e. lexical, linguistic, or structure heterogeneity, it is arduous to determine the outstanding alignments. But SNN-based OMT is able to obtain a high quality of alignments. As the real-world testing cases 3XX, although they have more complex heterogeneous characteristics than 1XX and 2XX, our proposal is an effective approach to solve the real-world OM problem and performs much better than other OMTs. To sum up, SNN-based OMT is able to effectively and efficiently solve the OM problem.

5 Conclusion

To solve the concept heterogeneity problem, in this paper, an SNN-based OMT is proposed, which is able to improve the matching efficiency by using a small part of RA to decrease the training time and improve the quality of matching result by using a logic reasoning approach to remove the conflict correspondences. The experimental results demonstrate that SNN-based OMT can determine high-quality alignment which outperforms the state-of-the-art OMTs. In the real matching scene, it is unrealistic to gain the RA. In the future, we will focus on the improvement of the efficiency and effectiveness of SNN-based OMT without using the RA, and be interested in capturing the semantic information by using both character-level and word-level information of concepts.

Acknowledgement. This work is supported by the Natural Science Foundation of Fujian Province (No. 2020J01875) and the National Natural Science Foundation of China (Nos. 61801527 and 61103143).

References

1. Antoniou, G., Van Harmelen, F.: A Semantic Web Primer. MIT Press, Cambridge (2004)
2. Bento, A., Zouaq, A., Gagnon, M.: Ontology matching using convolutional neural networks. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 5648–5653 (2020)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 34–43 (2001)
4. Chang, K.C., Chu, K.C., Wang, H.C., Lin, Y.C., Pan, J.S.: Energy saving technology of 5g base station based on internet of things collaborative control. IEEE Access **8**, 32935–32946 (2020)
5. Chen, C.H.: A cell probe-based method for vehicle speed estimation. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **103**(1), 265–267 (2020)
6. Chen, C.H., Song, F., Hwang, F.J., Wu, L.: A probability density function generator based on neural networks. Physica A Stat. Mech. Appl. **541**, 123344 (2020)
7. Chu, S.C., Dao, T.K., Pan, J.S., et al.: Identifying correctness data scheme for aggregating data in cluster heads of wireless sensor network based on Naive Bayes classification. EURASIP J. Wirel. Commun. Netw. **2020**(1), 1–15 (2020)
8. Da Silva, J., Revoredo, K., Baião, F.A., Euzenat, J.: Alin: improving interactive ontology matching by interactively revising mapping suggestions. Knowl. Eng. Rev. **35**, e1 (2020)

9. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Ontology matching: a machine learning approach. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. International Handbooks on Information Systems, pp. 385–403. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24750-0_19
10. Euzenat, J., Shvaiko, P.: Ontology Matching, vol. 18. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-49612-0
11. Jiang, C., Xue, X.: Matching biomedical ontologies with long short-term memory networks. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2484–2489. IEEE (2020)
12. Jiang, C., Xue, X.: A uniform compact genetic algorithm for matching bibliographic ontologies. Appl. Intell., 1–16 (2021)
13. Ali Khoudja, M., Fareh, M., Bouarfa, H.: A new supervised learning based ontology matching approach using neural networks. In: Rocha, Á., Serrhini, M. (eds.) EMENA-ISTL 2018. SIST, vol. 111, pp. 542–551. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03577-8_59
14. Lin, J.C.W., Shao, Y., Djenouri, Y., Yun, U.: ASRNN: a recurrent neural network with an attention model for sequence labeling. Knowl. Based Syst. **212**, 106548 (2021)
15. Lin, J.C.W., Shao, Y., Zhou, Y., Pirouz, M., Chen, H.C.: A Bi-LSTM mention hypergraph model with encoding schema for mention extraction. Eng. Appl. Artif. Intell. **85**, 175–181 (2019)
16. Liu, H., Wang, Y., Fan, N.: A hybrid deep grouping algorithm for large scale global optimization. IEEE Trans. Evol. Comput. **24**(6), 1112–1124 (2020)
17. Mao, M., Peng, Y., Spring, M.: Ontology mapping: as a binary classification problem. Concurr. Comput. Pract. Exp. **23**(9), 1010–1025 (2011)
18. McIlraith, S.A., Son, T.C., Zeng, H.: Semantic web services. IEEE Intell. Syst. **16**(2), 46–53 (2001)
19. Patel, A., Jain, S.: A partition based framework for large scale ontology matching. Recent Patents Eng. **14**(3), 488–501 (2020)
20. Rhayem, A., Mhiri, M.B.A., Gargouri, F.: Semantic web technologies for the internet of things: systematic literature review. Internet Things, 100206 (2020)
21. Xue, X., Wang, Y.: Optimizing ontology alignments through a memetic algorithm using both matchfmeasure and unanimous improvement ratio. Artif. Intell. **223**, 65–81 (2015)
22. Xue, X., Wang, Y.: Using memetic algorithm for instance coreference resolution. IEEE Trans. Knowl. Data Eng. **28**(2), 580–591 (2015)
23. Xue, X., Wu, X., Jiang, C., Mao, G., Zhu, H.: Integrating sensor ontologies with global and local alignment extractions. Wirel. Commun. Mob. Comput. **2021** (2021)
24. Xue, X., Yang, C., Jiang, C., Tsai, P.W., Mao, G., Zhu, H.: Optimizing ontology alignment through linkage learning on entity correspondences. Complexity **2021** (2021)
25. Xue, X., Yao, X.: Interactive ontology matching based on partial reference alignment. Appl. Soft Comput. **72**, 355–370 (2018)
26. Xue, X., Zhang, J.: Matching large-scale biomedical ontologies with central concept based partitioning algorithm and adaptive compact evolutionary algorithm. Appl. Soft Comput., 107343 (2021)