TOAST Results for OAEI 2012

Arkadiusz Jachnik, Andrzej Szwabe, Pawel Misiorek, and Przemyslaw Walkowiak

Institute of Control and Information Engineering, Poznan University of Technology,
M. Sklodowskiej-Curie Square 5, 60-965 Poznan, Poland
{arkadiusz.jachnik,andrzej.szwabe,pawel.misiorek,przemyslaw.walkowiak}@put.poznan.pl

Abstract. The Tensor-based Ontology Alignment SysTem (TOAST) is a general-purpose (i.e., domain-unspecific) self-configurable (i.e., requiring no user intervention) ontology matching tool. TOAST is based on one of the first tensor-based approaches to Statistical Relational Learning. Being one of the possible applications of the Statistical Relational Learning framework, TOAST may be seen as a system realizing a probabilistic inference with regard to a single relation only - the relation representing the 'semantic equivalence' of ontology classes or their properties. Due to the flexibility of the integrated tensor-based representation of heterogeneous data, TOAST is able to learn the semantics equivalence relation on the basis of partial matches data included in a train set.

1 Presentation of the System

The Tensor-based Ontology Alignment SysTem (TOAST) presented in this paper is an application of an extended version of our tensor-based approach to Statistical Relational Learning (SRL) referred to as Tensor-based Reflective Relational Learning Framework (TRRLF) [12]. In general, SRL is one of the most intensively investigated problems of Artificial Intelligence. Recently proposed tensor-based SRL methods are widely regarded (e.g., see [9]) as a promising alternative to the commonly used graphical models, such as Bayesian Networks and Markov Logic Networks [2], [5]. To our knowledge, TOAST represents the first tensor-based approach to ontology alignment.

We use a 3rd-order tensor as a data structure that is suitable to represent data provided as a set of RDF triples [4], [9]. There are several recent works considering the use of tensors to represent relational data given as RDF triples [5], [9], [4], [11]. The authors of these works assume that the *active* mode (corresponding to the RDF subject role) and the *passive* mode (corresponding to the RDF object role) of each entity have to be modeled as two separate tensor modes. However, they do not address the questions of (i) how to model the relation between two modes of the same entity and (ii) how the orientation of this relation (i.e., the setting which entity plays the active and which entity plays the passive role, as far as a given relation is concerned) influences the system performance [9], [4], [11].

We intend to confront these issues by proposing to model data in a way that enables a high level of flexibility for specifying the roles that any pair of entities plays with regard to any relation. Consequently, we represent both the *active* and *passive* modes of a given entity as potentially fully independent of each other – it is the correlation of the active mode and the passive mode (observable in the input data) that fully determines the extent to which the vectors representing the modes are algebraically similar to each other.

As we have shown in our experiments, the proposed tensor-based representation of relational data (in particular RDF triples), is appropriate for the ontology alignment task. It is worth noting that the internal data representation of TOAST is based on a probabilistic model of a vector space that has so far only been used in quantum Information Retrieval [13].

It should be stressed that TOAST does not require the use of external knowledge sources, such as dictionaries or thesauruses, in order to provide high quality results. However, the use of such knowledge data is possible – it may be realized by converting the data into the subject-predicate-object format [12], as discussed in Section 3.

1.1 State, Purpose, General Statement

TOAST is a fairly general-purpose ontology alignment tool. Being a specialized application of our SRL framework (i.e., the TRRL framework), TOAST may be seen as a system realizing a probabilistic inference with regard to a single relation only - the relation representing the semantic equivalence of ontology classes or their properties. The TRRL's flexibility, which is typical of SRL methods, is clearly visible in the propositional representation of all the heterogeneous data provided to the system (including the propositional representation of the occurrence of terms in the labels of the ontology classes).

The evaluation of TOAST has focused on the Anatomy track, which belongs to OAEI tracks that involve the use of the most expressive ontologies [3]. For this reason, we have not prepared the TOAST system to parse input data for any OAEI track other than the Anatomy. As a result, the Anatomy test is the only OAEI track test that TOAST passes. On the other hand, it should be noted that, in 2012, TOAST is the only matching system that can exploit additional partial alignments in the Anatomy track. To illustrate this fact, we present an additional experimental evaluation that has been performed, as suggested by the OAEI organizers, with the use of the OAEI 2010 dataset[1], in case of which the train set includes partial alignments. We show that, when partial alignments are available, TOAST is able to learn the semantics of all the relations [8], including the *matchesTo* relation, on the basis of the partial alignments data. It allows the system to exploit 'a behavioral dimension' of the alignments modeling and generation [12].

The results of TOAST evaluation presented in this paper are comparable with the results of the leading systems that have been evaluated from the perspective of Subtask #4 of the 2010 Anatomy track edition[2].

[1] Anatomy 2010 modified dataset: `http://oaei.ontologymatching.org/2010/anatomy/modifications2010.html`

[2] Anatomy - Results of 2010 Evaluation: `http://oaei.ontologymatching.org/2010/results/anatomy/index.html`

1.2 Specific Techniques Used

As TOAST is based on an SRL method, all techniques that are used in the system may be regarded as SRL solutions, rather than solutions specific to the ontology matching task. From such a general perspective, TOAST may be seen as a system that exploits a new algebraic data representation and processing method as a means for ontology alignment.

Tensor-Based Relational Data Representation

The tensor used in TOAST [12] can be seen as tensor product $T_{i,j,k} = [t_{i,j,k}]_{n \times n \times m} = S \times O \times R$ of vector spaces whose coordinates correspond to the set of subjects S, the set of objects O, and the set of relations R. We assume that $|R| = m$ and that $|S| = |O| = n$. Additionally, we define set F as a set of all the known facts (i.e., RDF triples) which are used to build the input tensor. The number $|F| = f$ determines the number of positive cells in the input tensor. Moreover, we define set $E = S \cup O \cup R$ as a set of elements (i.e., subjects, objects, and relations) used in the input data and represented in T by a slice (2nd-order array) of the 3rd-order tensor [12]. Due to the flexibility of the proposed tensor data model, it is possible to integrate the information about the ontology schema structure with the lexical knowledge. Therefore, set F contains facts about the relations between the ontology entities as well as between the ontology entities and the terms (representing lexical information) [12].

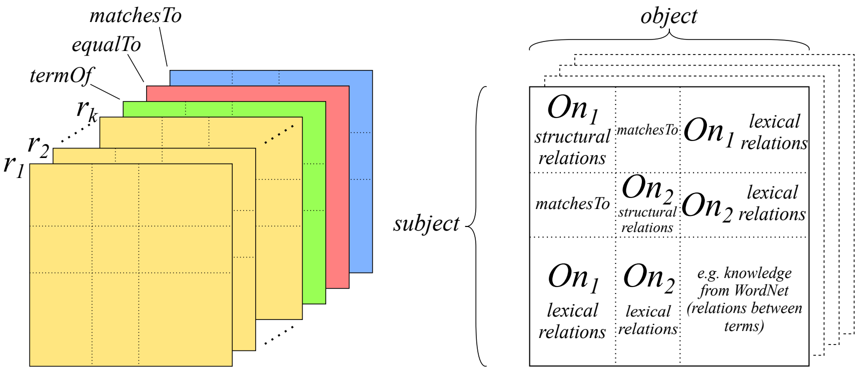

Fig. 1. The TRRLF tensor model and the TOAST tensor slice model.

Each tensor slice merges several submatrices and may be interpreted as a block matrix, as illustrated in Figure 1. For the case of a slice with structural information (i.e., representing *subClassOf* or *partOf* relations), the two submatrices on the diagonal represent a given relation for source ontology On_1 and target ontology On_2, respectively. Lexical relation slices (e.g., *termOf* slices) contain term-node submatrices describing the occurrences of terms in labels of the ontology classes. It should be stressed that TOAST allows us to use additional knowledge in the form of *partial reference alignments*. These partial alignments are represented by entries of an additional slice. Each of these entries represents the extent of the *matchesTo* relation between a given pair of nodes.

Common Vector Space

TRRLF uses a common d-dimensional vector space [12] to represent *context vectors* for all subjects from S, objects from O, and relations from R, as well as for all facts from F stored in the input tensor T. The context vectors set is modelled as matrix $X = [x_{i,j}]_{(2n+m+f)\times d}$, where:

$$X_{(2n+m+f)\times d} = \begin{bmatrix} X^E_{(2n+m)\times d} \\ X^F_{f\times d} \end{bmatrix}, \text{ where } X^E_{(2n+m)\times d} = \begin{bmatrix} X^S_{n\times d} \\ X^O_{n\times d} \\ X^R_{m\times d} \end{bmatrix}. \qquad (1)$$

The matrices X^E and X^F store context vectors of the elements from E and facts from F, respectively. X^E consists of three submatrices X^S, X^O, and X^R. The initial form of matrix X is prepared with the use of the random indexing procedure which ensures that non-zero values are uniformly distributed [1].

Learning and Matching Generation

In our approach, the learning procedure is based on the updating of the *context vectors*. The procedure is executed in steps, called *reflections*. A given reflection involves the reflective data processing [12], which is similar to Reflective Random Indexing [1]. As a result of modeling predicates as contex vectors, the system is able to process multi-relational data.

We introduce matrix $A = [a_{i,j}]_{(2n+m)\times f}$ as the source of data used in the learning process. Matrix A is constructed as a result of the 'flattening' operation applied to a tensor through all three dimensions (modes) [12].

The learning process consists of consecutive reflections. Each reflection consists of the training step (i.e., the context vector update based on learning matrix A) and the normalization step (based on the 3-norm) [12]. The method involves the application of the entropy-based criterion to indicate the optimal number of reflections. The description of this criterion is beyond the scope of this paper.

The matching likelihood prediction procedure is based on the use of the 1-norm of the Hadamard product of three vectors from X: vector $x^S_{i,\cdot}$ which corresponds to the ontology class in the subject mode, vector $x^O_{j,\cdot}$ which corresponds to the ontology class in the object mode and $x^R_{k,\cdot}$ which corresponds to the *matchesTo* relation. More formally, the probability that a match exists between the entities of the input ontologies is calculated according to the following formula:

$$p_{i,j,k} = \|x^S_{i,\cdot} \circ x^O_{j,\cdot} \circ x^R_{k,\cdot}\|_1 .$$

1.3 Adaptations Made for the OAEI Evaluation

For the OAEI Anatomy track evaluation, the TOAST input tensor has been generated using the following information extracted from the two input ontologies and partial reference alignments:

- structural information represented by relations *subClassOf* and *partOf*,
- lexical information represented by relation *hasTerm* and its inversion *termOf* (as explained above, we use both *hasTerm* and *termOf* relations, in order to avoid imposing an arbitrary direction of the lexical relation),
- lexical information represented by two additional slices built on the basis of `oboInOwl:hasRelatedSynonym` and `oboInOwl:hasDefinition`,
- additional partial reference alignments (i.e., the *matchesTo* relation) represented by an additional slice.

1.4 Link to the System and Parameters File

The TOAST system is available at `www.cie.put.poznan.pl/toast/TOAST_2012.zip`. The TOAST alignments (in the RDF alignment format) together with the configuration files for OAEI 2012 and OAEI 2010 may be found at `www.cie.put.poznan.pl/toast/results2012.zip`.

2 Results

In this section, we present the results of the evaluation of TOAST performed as part of the OAEI 2012 campaign. We have participated only in the Anatomy track of OAEI 2012. This year TOAST has been identified as the only matching tool evaluated in OAEI that is able to exploit partial alignments of the Anatomy track. Unfortunately, for this reason the organizers have dropped this specific type of evaluation. Nevertheless, we have decided to show that our system is able to effectively use the additional partial alignments from the OAEI 2010 edition dataset (see Subtask #4 in the OAEI 2010 edition).

The official OAEI evaluation procedure has been executed on an Ubuntu machine with 2-core x64 processor and 4GB RAM. We additionally present the results obtained when using our machine with Ubuntu OS, 4-core processor and 16GB RAM.

2.1 Anatomy 2012 Track

OAEI 2012 Evaluation Table 1 gathers the results of the TOAST system evaluation expressed in terms of precision (P), recall (R), the F_1 measure, the number of returned matches (RM), true positives (TP), false positives (FP), false negatives (FN), trivial true positives (TP-trivial), and non-trivial true positives (TP-non-trivial). Two experiments have been executed: one by the organizers of the OAEI campaign and the other by the authors.

Table 1. The results of TOAST evaluation in the Anatomy 2012 track.

No.	TOAST config.:	P	R	F_1	RM	TP	FP	FN	TP-trivial	TP-non-trivial	time $[s]$
1	OAEI official evaluation	0.854	0.755	0.801	*no data available*						3464[1]
2	Auto-config	0.852	0.749	0.797	1333	1136	197	380	914	222	1218[2]

[1] execution time on the OAEI organizers' machine,
[2] execution time on the authors' machine.

Anatomy 2010 with Additional Partial Alignments Table 2 presents the results of our system application in the Anatomy Subtask #4 track involving the use of partial alignments. The experiments show TOAST operating in its default, fully automatic mode. Besides using the additional knowledge derived from partial alignments, we have also used information about synonyms embedded in the ontologies (relation `oboInOwl:hasRelatedSynonym`). This information has been stored as an additional tensor slice with the lexical data.

Table 2. The results of TOAST evaluation for the Anatomy 2010 Subtask #4 dataset.

No.	TOAST config.:	P	R	F_1	RM	TP	FP	FN	TP-trivial	TP-non-trivial	time[1] $[s]$
3	Auto-config	0.885	0.776	0.827	1332	1179	153	341	930	249	1941
4	Auto-config + synonyms	0.908	0.789	0.844	1320	1199	121	321	932	267	2476

[1] execution time measured on the authors' machine.

2.2 Benchmark, Conference, Multifarm, Library, Large Biomedical Ontologies, and Instance Matching tracks

As already mentioned above, in our research on TOAST, we have focused only on the Anatomy track. For this reason, we have not prepared our tool to parse input data for any OAEI track other than Anatomy, namely Benchmark, Conference, Multifarm, Library, Large Biomedical Ontologies, and Instance Matching tracks.

3 General Comments

In the following section, we provide the general comments about the TOAST results and future improvements as well as our suggestions concerning possible directions of the OAEI contest enhancements.

3.1 Comments on the Results

In the OAEI Anatomy 2012 evaluation, the self-configuring variant of TOAST achieves a comparatively high quality (see Table 1). The slight difference between our results and the results obtained by the organizers is due to the application of slightly different techniques for computing precision and recall. In our experiment, we have used the standard method that is featured by the SEALS client.

On the basis of the results of the OAEI Anatomy 2010 evaluation, it may be additionally demonstrated that the availability of partial alignments allows TOAST to improve the matches quality. Moreover, evaluation 4 (see Table 2) has revealed the importance of additional lexical knowledge (synonyms) for improving the quality of TOAST-generated mappings.

3.2 Discussions on the Way to Improve the Proposed System

The TOAST version prepared for OAEI does not use any domain-specific background knowledge sources (such as biomedical ontologies). However, the relational data representation and processing capabilities of TOAST enable the system to exploit any generic knowledge source or linguistic resource such as WordNet [10]. In the case of any SRL-based ontology matching system (such as TOAST), taking the advantage of using external data sources (especially sources of structured data) is comparatively easy.

3.3 Comments on the OAEI Anatomy Dataset with the Partial Alignments

We have performed a lexical analysis of the Anatomy dataset and have identified several subsets of different types of alignments present in this dataset. As a result of this analysis, it has been established that the set of reference alignments contains 933 *trivial* matches (i.e., literal matches that can be found by simple string comparison) and 587 *non-trivial* matches (that require more sophisticated analysis to be identified), while the set of partial matches consists of 928 *trivial* matches and only 59 *non-trivial* matches.

This shows that the set of partial alignments contains $\sim 65\%$ of the reference matches set. However, the analogical proportion of the *trivial* and *non-trivial* matches numbers differs greatly. It can be concluded that Anatomy 2010 dataset including partial alignments is rather poorly balanced, which makes it unsuitable for a reliable evaluation in relevance feedback scenarios. Following the methodology widely used in the field of Information Retrieval [6], we suggest the development of a new subset of partial matches randomly chosen from the set of reference matches. We believe that such a dataset modification will help to increase the interest in Anatomy Subtask #4, which is the only OAEI scenario that deals with the use of relevance feedback data.

3.4 Proposed New Measures

We suggest the extension of the evaluation measures set by the Area Underneath an ROC curve - AUROC measure [6]. AUROC is a widely-used probabilistically interpretable classification quality measure. Although using AUROC requires the availability of data on incorrect matches, unknown mappings do not influence the AUROC measurement result. AUROC is regarded as the best recommendation quality measure, as long as one assumes that the purpose of an evaluated recommender is to sort all items according to their estimated usefulness. Therefore, the AUROC results may enrich the

OAEI evaluation by enabling not only the examination of the matching results given as a set, but also the evaluation of the order of the matches generated by means of the evaluated systems.

4 Conclusions

TOAST, as an application of the general-purpose Tensor-based Reflective Relational Learning framework, may be regarded as a universal (i.e., domain-unspecific) ontology matching tool. To our knowledge, TOAST is the first tensor-based approach to ontology alignment that integrates the structural and lexical data in a relational way. We have shown that the system is self-configurable and provides high-quality results. Moreover, the tool is able to effectively use partial matches data.

Acknowledgments. This work is supported by the Polish Ministry of Science and Higher Education under grant N N516 196737, and by Poznan University of Technology under grant DS 45-085/12.

References

1. Cohen, T., Schaneveldt, R., Widdows, D.: Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. Journal of Biomedical Informatics 43(2), 240256 (2010)
2. De Raedt, L.: Logical and Relational Learning. Springer (2008)
3. Euzenat, J., Ferrara, A., van Hage, W. R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Svb-Zamazal, O., dos Santos, C.T.: Results of the ontology alignment evaluation initiative 2011. In: OM-2011 (2011)
4. Franz, T., Schultz, A., Sizov, S., Staab, S.: Triplerank: Ranking Semantic Web Data by Tensor Decomposition. In: The Semantic Web-ISWC 2009, pp. 213–228 (2009)
5. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. MIT Press (2007)
6. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. ACM Trans. Information Systems, 22(1), 5–53 (2004)
7. Kolda, T. G., and Bader, B. W. Tensor Decompositions and Applications, SIAM Review 51(3):455-500 (2009)
8. Lavrenko V., A Generative Theory of Relevance. Springer-Verlag, Berlin (2010)
9. Nickel, M., Tresp, V., Kriegel, H.P.: A Three-Way Model for Collective Learning on Multi-Relational Data. In Proceedings of the 28th International Conference on Machine Learning, pp.809-816 (2011)
10. Princeton University. WordNet - A lexical database for English. Online: `http://wordnet.princeton.edu`
11. Sutskever, I., Salakhutdinov, R., Tenenbaum, J. B.: Modelling Relational Data Using Bayesian Clustered Tensor Factorization. Advances in Neural Information Processing Systems, 22 (2009)
12. Szwabe, A., Misiorek, P., Walkowiak, P.: Tensor-based Relational Learning for Ontology Matching. In Grana M. et al. (eds.) Advances in Knowledge-Based and Intelligent Information and Engineering Systems (KES2012), Frontiers in Artificial Intelligence and Applications vol. 243, pp. 509-518. IOS Press (2012)
13. van Rijsbergen, C. J.: The Geometry of Information Retrieval. Cambridge University Press. New York, USA (2004)