Edinburgh Associative Thesaurus as RDF and DBpedia Mapping

Jörn Hees, Rouven Bauer, Joachim Folz, Damian Borth, and Andreas Dengel

[1] Computer Science Department, University of Kaiserslautern, Germany
[2] Knowledge Management Department, DFKI GmbH, Kaiserslautern, Germany
{firstname.lastname}@dfki.de

Abstract. Associations, which are one of the key ingredients of human intelligence and thinking, are not easily accessible to the Semantic Web community. High quality RDF datasets of this kind are missing. In this paper we generate such a dataset by transforming 788 K free-text associations of the Edinburgh Associative Thesaurus (EAT) into RDF. Furthermore, we provide a verified mapping of strong textual associations from EAT to DBpedia Entities with the help of a semi-automatic mapping approach. Both generated datasets are made publicly available and can be used as a benchmark for cross-type link prediction and pattern learning.

1 Introduction

Associations as one of the building blocks of human intelligence, thinking, context forming and everyday communication [4] are not well represented in currently published Linked Data datasets. This impedes AI research: due to the missing ground truth of semantic entities which are associated by humans, we can neither analyse human associations in existing datasets, nor train machines to learn graph patterns for them.

2 Related Work

Previously, we developed semantic games with a purpose to collect a semantic association ground truth (Linked Data Games [7], KnowledgeTestGame [5]) or to rank existing triples by association strengths (BetterRelations [6]). Along the lines of fact ranking ground truth datasets, other works such as WhoKnows [9] and more recently FRanCo [2] have been published. While fact ranking in general only focuses on existing facts, FRanCo in its first step also collected free-text fact input about the entity in question, resulting in $\sim$ 7.8 K raw free-text facts and a NER mapping back to semantic entities[3]. While these works can help collecting new associations, the datasets generated in this paper are orders of magnitude larger, published in RDF, and each of their mappings has been manually verified in order to provide high precision ground truth for machine learning.

[3] http://s16a.org/node/13

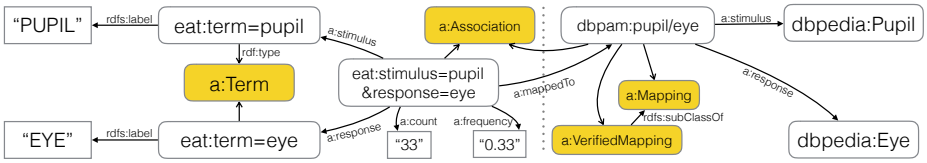

Fig. 1. Example of the EAT associations "pupil - eye" as RDF (left) and its mapping to the semantic association `dbpam:pupil/eye` between the DBpedia entities `dbpedia:Pupil` and `dbpedia:Eye` (right).

Prefixes: `a:` <`https://w3id.org/associations/vocab#`>
`eat:` <`http://www.eat.rl.ac.uk/#`>
`dbpedia:` <`http://dbpedia.org/resource/`>
`dbpam:` <`https://w3id.org/associations/mapping_eat_dbpedia#`>

3 Edinburgh Associative Thesaurus as RDF

EAT [8] was created in the 1970s and is a dataset of single free-text associations collected directly from humans. It consists of a well connected network of $\sim$ 788 K *raw associations* which form $\sim$ 326 K *unique associations* (unique stimulus-response-pairs) between 8200 unique stimuli and $\sim$ 22700 unique responses.

About 5000 unique associations occur more than 20 times (167 K raw associations). In the remainder of this paper we will refer to them as *strong associations*. An example for such a strong association is the one between stimulus "dog" and response "cat" which occurred 57 out of 100 times.

As the EAT dataset[4] is not available as RDF, we create an association vocabulary[5] and use it to transform EAT into RDF (see example in Figure 1). We formally model EAT as a multi-set of raw associations. Each raw association $a \in EAT$ is a free-text stimulus-response-pair: $a = (s, r), s \in S, r \in R$. The union of all stimuli S and responses R forms the set of terms $T = S \cup R$. Further, we can define the count $c_{s,r}$ as the number of occurrences of the raw association and the relative frequency $f_{s,r}$ as the relative count of response r with respect to a fixed stimulus s over all responses to that stimulus. The resulting transformation of EAT into RDF consists of 1 674 376 triples[6].

4 Mapping EAT to DBpedia

This section describes the process of mapping associations from EAT to equivalent semantic associations between pairs of DBpedia [1] entities. If we find such two entities, we call the relation between them a *semantic association*.

For example, let's focus on the association "pupil - eye", with URI `eat:stimulus=pupil&response=eye` in Figure 1. We can identify two DBpedia entities, namely `dbpedia:Pupil` and `dbpedia:Eye` with the intended meaning of the

[4] `http://www.eat.rl.ac.uk/`
[5] `https://w3id.org/associations/vocab#`
[6] `https://w3id.org/associations/eat.nt.gz`

association and create a new semantic association `dbpam:pupil/eye` with the corresponding links.

For the mapping we focused on the $\sim$ 5000 unique strong associations occurring more than 20 times (167 K raw associations), as they are more robust with respect to subjectivity, location and time dependency.

Based on experiences gained from a manual mapping of a random sample[7], we were able to develop an automatic mapping approach with the following scoring component (non-exclusive likelihoods and examples in brackets) which uses the Wikipedia API[8]:

- **Composite phrases** (28 %, e.g., "port - wine"): As a composite phrase is a name for a single semantic entity it is a bad candidate for a semantic association (between two different semantic entities). Hence, if searching for Wikipedia articles (or redirect pages) containing stimulus and response in their title is successful, the mapping's score receives a strong punishment.
- **Adjectives & verbs vs. nouns** (22 %, e.g., "unbound - free"): Due to Wikipedia's nature of being an encyclopaedia, adjectives and verbs are underrepresented in contrast to nouns. To identify such cases, the stimulus and response are searched in Wordnet [3], potentially resulting in multiple synset candidates for each. Mappings containing only synset candidates with the given type "noun" are preferred. The more synset candidates with types unequal to "noun" are found, the stronger the punishment for the mapping's score.
- **Reflexive mappings / synonyms** (18 %, e.g., "children - kids"): If the mapping of both the stimulus and the response result in the same semantic entity, the score is strongly punished.
- **Plural words** (16 %, e.g., "thumbs - fingers"): A simple stemming approach is used to compare the stimulus/response to the identified Wikipedia article titles after following redirects. If the match is close to perfect and only differs in singular/plural, the score only receives a slight punishment.
- **Disambiguation pages** (16 %, e.g., "pod - pea"): If the mappings of stimulus or response result in a Wikipedia disambiguation page, the mapping's score receives a strong punishment.

After applying the automatic mapping to the $\sim$ 5000 strong associations, the top scoring 1066 semantic association candidates (corresponding to $\sim$ 34.2 K raw associations) were selected for human verification.

In order to quickly verify the 1066 mapping candidates, a small web application was used, which shows the textual association from EAT on top (stimulus - response) and the abstracts of both mapped Wikipedia articles below and asks the user if both stimulus and response are correctly mapped.

[7] The manual mapping showed that about 12 - 28 % of the 5000 strong associations are mappable to DBpedia entities (depending on the amount of human labour and intelligence involved).

[8] `http://www.mediawiki.org/wiki/API:Main_page`

The web application was used by 10 reviewers and allowed the verification (3 independent "Yes" ratings) of 790 of 1066 mappings (corresponding to $\sim$ 25.5 K raw associations).

For each of the 790 verified mapped semantic associations a mapping URI is created analogously to Figure 1. The resulting mapping dataset consisting of 4740 triples can be downloaded[9] or simply dereferenced.

5 Conclusion & Outlook

In this paper we presented a transformation of 788 K free-text associations from the Edinburgh Associative Thesaurus into a RDF dataset. Further, we presented a first mapping of its strong associations to semantic associations between DBpedia entities, resulting in 790 manually verified mappings corresponding to $\sim$ 25.5 K raw associations.

In the future we plan to conduct pattern learning based on the mapped semantic associations. As all generated datasets are publicly available, we also look forward to them being used as benchmark or ground truth datasets, for example for link prediction tasks.

This work was financed by the University of Kaiserslautern PhD scholarship program and the BMBF project MOM (Grant 01IW15002).

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009)
2. Bobić, T., Waitelonis, J., Sack, H.: FRanCo – A Ground Truth Corpus for Fact Ranking Evaluation. In: SumPre 2015 at ESWC 2015. pp. 1–12
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
4. Gerrig, R.J., Zimbardo, P.G.: Psychology and Life. Allyn & Bacon, Pearson, Boston, USA, 19th edn. (2010)
5. Hees, J., Khamis, M., Biedert, R., Abdennadher, S., Dengel, A.: Collecting Links between Entities Ranked by Human Association Strengths. In: ESWC. vol. 7882, pp. 517–531. Springer LNCS, Montpellier, France (2013)
6. Hees, J., Roth-berghofer, T., Biedert, R., Adrian, B., Dengel, A.: BetterRelations: Using a Game to Rate Linked Data Triples. In: KI 2011: Advances in Artificial Intelligence. pp. 134–138. Springer Berlin / Heidelberg, Berlin (2011)
7. Hees, J., Roth-Berghofer, T., Dengel, A.: Linked Data Games: Simulating Human Association with Linked Data. In: LWA 2010. pp. 255–260. Kassel, Germany (2010)
8. Kiss, G.R., Armstrong, C., Milroy, R., Piper, J.: An associative thesaurus of English and its computer analysis. In: The Computer and Literary Studies, pp. 153–165. Edinburgh University Press, Edinburgh, UK (1973)
9. Kny, E., Kölle, S., Töpper, G., Wittmers, E.: WhoKnows? (2010)

[9] `https://w3id.org/associations/mapping_eat_dbpedia.nt.gz`