Relevance-Based Evaluation of Alignment Approaches: the OAEI 2007 food task revisited

Willem Robert van Hage[1,2], Hap Kolb[1], and Guus Schreiber[2]

[1] TNO Science & Industry, Stieltjesweg 1, 2628CK Delft, the Netherlands,
hap.kolb@tno.nl
[2] Vrije Universiteit Amsterdam, de Boelelaan 1081a, 1081HV Amsterdam, the Netherlands,
wrvhage@few.vu.nl, schreiber@cs.vu.nl

Abstract. Current state-of-the-art ontology-alignment evaluation methods are based on the assumption that alignment relations come in two flavors: correct and incorrect. Some alignment systems find more correct mappings than others and hence, by this assumption, they perform better. In practical applications however, it does not only matter *how many* correct mappings you find, but also *which* correct mappings you find. This means that, apart from correctness, relevance should also be included in the evaluation procedure. In this paper we expand the sample-based evaluation of the OAEI 2007 *food task* with a sample evaluation that uses relevance to prototypical search tasks as a selection criterion for the drawing of sample mappings.

1 Introduction

In recent years ontology alignment has become a major field of research [3, 5]. Especially in the field of digital libraries it has had a great impact. Good evaluation is essential for the deployment of ontology-alignment techniques in practice. The main contribution of this paper is to offer a simple method to capture the performance of alignment approaches in actual applications. We introduce *relevance-based evaluation*, which compensates for some of the shortcomings of existing methods by using the needs of users during sample selection. We apply this method to the data of the OAEI 2007 *food task* [2].

Nearly all existing evaluation measures used to determine the quality of alignment approaches are based on counting mappings [1, 2]. For instance, in the context of ontology alignment, the definition of Recall is defined as the number of correct mappings a system produces divided by the total number of correct mappings that can possibly be found (*i.e.* that are desired to be part of the result). Regardless of their differences, most of these measures have one thing in common: They do not favor one mapping over the other in order to give an objective impression of system performance. Any mapping could prove to be important to some application. Therefore, they can only tell us *how many* mappings are found on average by a system, but not *which* mappings are found and whether the mappings that are found are those that are useful for a certain application. Whenever someone wants to decide which alignment approach is best suited for his application (*e.g.* [7]) he will have to reinterpret average expected performance in the light of his own needs. This can be a serious obstacle for users.

A solution to this problem is to incorporate the importance of mappings (*i.e.* relevance) into the evaluation result. This solution immediately raises two new problems: (1) How to come up with suitable importance weights, and (2) How to define a simple and intuitive way to use these weights With respect to problem 1, there are many sensible ways to weigh the importance of mappings. For example, based on the size of the logical consequence, *cf.* Semantic Precision and Semantic Recall [1], or on expected traversal frequency, *cf.* [4]. Relevance-based evaluation equates importance to relevance to prototypical application scenarios. Likewise, with respect to problem 2, there are many sensible ways to incorporate mapping importance into an evaluation method. For example, linear combination, *cf.* [6], or stratification *cf.* [9]. As opposed to existing methods to account for the relevance of mappings that include it as a variable in an evaluation measure, we use relevance to steer the sample-selection process. Instead of randomly selecting mappings for the evaluation of alignment approaches (*cf.* the *food* and *environment tasks* described in [2]) we select *only* those that are relevant to an application. This way we can use existing and well-understood evaluation metrics, like Precision and Recall, to measure performance on important tasks as opposed to fictive average-case performance.

2 Experimental Set-up

We demonstrate how relevance-based evaluation works by extending the existing results of the OAEI 2007 *food task*, which did not take relevance into account. We determine relevance for the mappings based on hot topics related to this task, like global warming and increasing food prices, which we obtain by means of query-log analysis, expert interviews, and news feeds. For the original OAEI 2007 *food task*, Recall was measured on samples that represent the frequency of topics in the vocabularies. In relevance-based evaluation the samples are drawn by the frequency of use in search tasks, specifically, finding documents about prototypical agricultural topics of current interest in one collection using the indexing vocabulary of the other. The procedure we use is as follows: (1) Gather topics that represent important use cases. We gather "hot" topics in agriculture from the query log files of the FAO AGRIS/CARIS search engine, the FAO newsroom website, and interviews with two experts. Patricia Merrikin from the FAO's David Lubin library, and Fred van de Brug, from the TNO Quality of Life food-safety group. We manually construct search-engine queries for each topic. (2) Gather documents that are highly relevant to these topics. We ascertain which documents would be sufficient for the hot topics by gathering suitable candidate documents from the part of the FAO AGRIS/CARIS and USDA AGRICOLA reference databases that overlaps. We use a free-text search engine[3] and manually filter out all irrelevant documents. (3) Collect the meta-data describing the subject of these documents and align the concepts that describe the subject of the documents to concepts in the other thesaurus. We collect values of the Dublin Core subject field from the AGRIS/CARIS and AGRICOLA reference databases. These values come from subject vocabularies, respectively AGROVOC and the NAL Agricultural Thesaurus. We manually align each concept to the most similar concept in the other vocabulary. The resulting mappings make up our sample set

[3] `http://www.fao.org/agris/search`

of relevant mappings. (4) Apply the mappings for evaluation by counting how many of these mappings have been found by ontology alignment systems and comparing system performance based on these counts. Specifically, we re-calculate Recall for the top-4 systems of the OAEI 2007 *food task*, following the same procedure as described in [2, 9], but use the new set of relevant mappings.

3 Sample Construction

Topics In order to get a broad overview of current affairs in the agricultural domain we gathered topics from three sources: Analysis of the search log files of the AGRIS/CARIS search engine, topics in the "Focus on the issues" section of the FAO Newsroom, and expert interviews. Detailed descriptions of the topics can be found at `http://www.few.vu.nl/~wrvhage/om2008/topics.html`.

Documents Per topic did a full-text search on the AGRIS/CARIS search engine limited to the set of documents that is shared between the AGRICOLA and AGRIS/CARIS collections and fetched the top-100 of the results. From these 1500 documents we selected only the ones that are relevant to our topics, on average 31 per query, and that have been assigned Dublin Core subject terms in both collections. This left 52 documents in total, on average 3.8 per query. For four of the topics we found no documents that were both relevant and indexed in both collections. The reason for this is that these topics are all very new issues. The greatest overlap between the AGRIS/CARIS and AGRICOLA collections exists for documents published between 1985 and 1995. After the year 2000 no documents have been imported and thus it is hard to find relevant documents for new issues. We assume that the 52 double-annotated relevant documents are representative of the set of all relevant documents with subject meta-data, *i.e.* also the documents with only annotations in one of the two collections. These are the documents for which alignment could make the biggest difference. This is a reasonable assumption, because the indexing process of both collections is regulated by a protocol to control continuity.

Mappings Having established which documents are potentially important to find, we have to decide which mappings will be of most benefit to someone who wants to find them. We assume that the mappings that map the subject annotations as strictly as possible to the other vocabulary are the most beneficial for any search strategy that employs them. Given this assumption, we manually constructed the set of mappings that connect each concept used to index the 53 relevant documents with its most similar counterpart.

The alignment of the 266 NALT concepts and 212 AGROVOC concepts was done by thesaurus experts at the FAO and USDA, Gudrun Johannsen and Lori Finch. This led to a sample reference alignment consisting of 347 mappings: 74 broadMatch / narrowMatch and 273 exactMatch (79%). 11 concepts had no exact, broader or narrower counterpart. This is a higher percentage of exactMatch mappings than we expected based on our experiences with the OAEI *food task*. For the *food task*, arbitrary subhierarchies of the AGROVOC and NAL thesaurus were drawn and manually aligned with the other thesaurus. Most of the resulting mappings were equivalence relations. The sample sets, the percentage of equivalence mappings in the reference alignment (*i.e.* the desired equivalence relations) varied between 54% and 71%.

4 Sample Evaluation Results

Having constructed a new sample reference alignment we can use it to measure the performance of alignment approaches. The measurement of Recall under the open-world assumption is inherently hard, so we choose to reiterate the evaluation of Recall on the OAEI 2007 *food task*. This gives us a second opinion on the existing evaluation. For the sake of simplicity we calculate Recall scores of the top-4 of the systems that participated in the OAEI 2007 *food task*. The results are shown in table 1.

	Falcon-AO	RiMOM	DSSim	X-SOM
OAEI 2007 food, only exactMatch (54% of total)	0.90	0.77	0.37	0.11
hot topics, only exactMatch (79% of total)	0.96 ↑	0.60 ↓	0.16 ↓	0.07 ↓
OAEI 2007 food, exact, broad, narrowMatch	0.49	0.42	0.20	0.06
hot topics, exact, broad, narrowMatch	0.75 ↑	0.47 ↑	0.12 ↓	0.05 ≈

Table 1. Recall of alignment approaches measured on sample mappings biased towards relevance to hot topics in agriculture and on impartial, non-relevance-based sample mappings from the OAEI 2007 *food task*. Arrows indicate significant differences (using the tests described in [9]).

There are a number of striking points to note about these results. For most systems there is a significant positive or negative difference. Overall, the difference with non-relevance-based evaluation is large. For exactMatch relations performance in general is lower for relevance-based evaluation than for non-relevance-based evaluation, with the exception of Falcon-AO, although the relative difference is small. However, the ranking of the alignment approaches is left unchanged. The results of relevance-based evaluation seem to exaggerate the differences between the performance of the approaches. This can be explained by the relatively high number of obvious matches (93%) in the set of mappings on hot topics. None of the approaches was able to find a substantial number of difficult mappings, but the best approaches were good at finding all obvious mappings before resorting to speculation about the harder mappings. The best two systems, Falcon-AO and RiMOM performed relatively good when accounting for all relation types, the last row of table 1, even though they found no broadMatch and narrowMatch relations. This is due to the kind of exactMatch relations they *did*, which were mostly of the obvious kind (*i.e.* literal matches), which was exactly the kind that was needed most for the hot topics. The high percentage of exactMatch relations in the set on hot topics accentuates their behavior. The converse goes for DSSim, which found a relatively low number of obvious mappings. Fewer broadMatch and narrowMatch mappings seem to be needed than one would expect from the non-relevance-based evaluation method. Compare the percentage in the OAEI 2007 Recall set, 54%, to the percentage based on hot topics, 78.6%. Although there is a large part of the AGROVOC and NALT vocabularies that does not have a counterpart in the other vocabulary, the portion that is actually used suffers less than one would expect from this mismatch. Apparently, indexers mainly pick their terms from a limited set, which shows a greater overlap. (After all, why needlessly complicate things?) It remains to be seen if this also applies to other vocabulary mappings. On one hand this means that approaches that can only find equivalence mappings perform better in practice than was expected. On the other hand it confirms the expectation that a large part (*more than 20%*) of the mappings that are needed for federated search over AGRIS/CARIS and AGRICOLA consists of other relations than equivalence relations. Also, one can conclude that systems that are incapable of finding a substantial number of equivalence relations can only play a marginal role.

5 Discussion

By using relevance as a sample criterion we avoid having to come up with an artificial approximation of importance. We can simply explore the performance difference on samples consisting of relevant mappings and samples consisting of irrelevant mappings. Under minimal assumptions we avoid having to choose a specific retrieval method while retaining the the character of an end-to-end evaluation. (*cf.* the *End-to-end Evaluation* method described in [9]) This saves us the effort of extensive user studies while not ignoring the behavior of alignment approaches in real-life situations. Considering the fact that AGROVOC and NALT are two of the most widely used agricultural ontologies, and that they are prototypical examples of domain thesauri in their design we conclude the following. From the point of view of a developer of a federated search engine in the agricultural domain that needs an alignment we can conclude that at the moment the Falcon-AO is a good starting point. In the case described in this paper, Falcon-AO found three quarters of the mappings. This empirical study has shown that at least 20% of the required mappings to solve the typical federated-search problem described in this paper are hierarchical relations. Even though this is a smaller fraction than we initially expected it is still a large part. An extended version of this paper can be found in [8].

References

1. Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proc. of IJCAI 2007*, pages 348–353, 2007.
2. Jérôme Euzenat, Malgorzata Mochol, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb, Vojtěch Svátek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative, 2007.
3. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007. ISBN 978-3-540-49611-3.
4. Laura Hollink, Mark van Assem, Shenghui Wang, Antoine Isaac, and Guus Schreiber. Two variations on ontology alignment evaluation: Methodological issues. In *Proc. of ESWC*, 2008.
5. Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The knowledge engineering review*, 18(1):1–31, march 2003.
6. Jaana Kekäläinen. Binary and graded relevance in ir evaluations–comparison of the effects on ranking of ir systems. *Information Processing and Management*, 41(5):1019–1033, 2005.
7. Malgorzata Mochol, Anja Jentzsch, and Jérôme Euzenat. Applying an analytic method for matching approach selection. In *Proc. of OM-2006*, pages 37–48, 2006.
8. Willem Robert van Hage. *Evaluating Ontology-Alignment Techniques*. PhD thesis, Vrije Universiteit Amsterdam, 2008. http://www.few.vu.nl/∼wrvhage/thesis.pdf.
9. Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski. Sample evaluation of ontology-matching systems. In *Proc. of EON*, 2007.