Combining Vocabulary Alignment Techniques

Anna Tordai
VU University Amsterdam
atordai@cs.vu.nl

Jacco van Ossenbruggen
CWI Amsterdam
jrvosse@cs.vu.nl

Guus Schreiber
VU University Amsterdam
schreiber@cs.vu.nl

ABSTRACT
Identifying alignments between vocabularies has become a central knowledge engineering activity. A plethora of alignment techniques has been developed over the past years. In this paper we present a case study in which we examine and evaluate the practical use of three typical alignment techniques. The study involves the alignment of two vocabularies used in a semantic-search engine for cultural-heritage objects. We show that a sequence can be beneficial. The case study gives insight into evaluation issues, such as techniques for identification of false positives. We see this work as a step to a badly-needed methodology for alignment.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*semantic networks*

Keywords
vocabulary alignment, case study, cultural heritage

1. INTRODUCTION
In the past years there has been tremendous activity in the ontology alignment field. A large number of techniques and algorithms have been developed [6, 4]. Within the Ontology Alignment Evaluation Initiative (OAEI[1]) alignment techniques are applied to benchmark data. However, despite these efforts there is still a clear lack of methodological support for selecting an appropriate (subset of) alignment technique(s). This paper presents a case study in ontology alignment. The application context is the MultimediaN E-Culture project [10]. This project deploys a large number of vocabularies of different heritage collections. These include large vocabularies such as the Getty theasauri and WordNets for different languages, but also smaller collection-specific vocabularies. Readers unfamiliar with the application context may want to have a look at the general E-Culture demonstrator[2] or the Europeana Thought Lab demonstrator[3]. Vocabulary alignments are an crucial element of the semantic interoperability realized by these systems.

In this paper we investigate the application of three alignment techniques to two vocabularies from the E-Culture repository. The general objective of the case study is to gain insight into methodological issues related to alignment-technique selection. In particular, we are interested in the following two research questions:

1. *Can we show added value of combined use of different alignment techniques?* The OAEI study has shown that the performance of techniques is dependent on the application context and that no single group of techniques can be identified as being superior. Therefore, combining techniques appears to be the obvious way to go, in particular to increase recall.

2. *Can we improve alignment precision by deploying techniques for identifying false positives?* Higher recall is likely to lead to lower precision. In the paper we examine techniques for pruning the set of candidate alignments and identify likely false positives.

This paper is structured as follows: In Section 2 we discuss related work relevant for methodological issues of ontology alignment. Section 3 describes the setup of the case study. Section 4 describes the results of the combined application of alignment techniques. In Section 5 we look at techniques for improving precision. Section 6 discusses what we have learnt with respect to a

[1] `http://oaei.ontologymatching.org/`

[2] `http://e-culture.multimedian.nl/demo/session/search`

[3] `http://eculture.cs.vu.nl/europeana/session/search`

future alignment methodology. We postulate a number of potential avenues forward.

2. RELATED WORK

There is comparatively little work on procedures and guidelines in ontology alignment. Euzenat et al.[4] identify application requirements and propose a case-based method for recommending alignment techniques, in which application dimensions are correlated with properties of alignment tools to determine the best fit. This work is based on outcomes of the respective OAEI studies. One tool, RiMOM[12], closely followed by Falcon[7] had the best fit for each application. The applications themselves are highly abstract, such as "schema integration" and "multi agent communication".

Aleksovski et al. [1] performed a survey of techniques for alignment problems and based the alignment cases on existing ontology alignment applications. They list several applications, among these are STITCH Cultural Heritage browser[4], The Unified Medical Language System (UMLS) and Internet Directories. The applications and their more abstract types are categorized according to priorities regarding the quality of alignment and the complexity of representation of the ontologies. Generalizing from the techniques used in the applications, they propose techniques for each alignment type. For example, STITCH is typed as "unified view over collections" and the suggestion is that lexical alignment can solve a large part of the alignment problem for other applications that can be typed similarly.

The OAEI workshops' aim is the comparison of ontology matching tools on predefined test sets. The tools use various techniques and their combinations for performing ontology alignments. An overview of alignment techniques can be found in Euzenat and Shvaiko [6]. Tools implementing alignment techniques or their combinations take part in a number of tracks. These include a "benchmark track", a "expressive ontologies track", and a "directories and thesauri track". The challenges for the tools vary depending on the track. For example, one task in the expressive ontologies track is to align anatomical ontologies which are complex and use specialized vocabularies. The vocabularies of the library task in the directories and thesauri track contain less structural information and are in Dutch language. Although not all systems participating in the OAEI take part in each track, certain systems tend to perform better than others.

In the 2007 OAEI workshop [5], Falcon stood out with a consistently good performance across most tracks. Falcon uses a combination of lexical comparison and statistic analysis with structural similarity techniques and graph partitioning. In the 2008 OAEI workshop [2], where Falcon no longer took part, the top performing systems such as RiMOM and DSSim[9] all use combinations of techniques for generating alignments.

[4] http://www.cs.vu.nl/STITCH/

3. CASE STUDY SETUP

Data sets

In the E-Culture project we have a large number of small collection-specific vocabularies and a few large general purpose vocabularies. Aligning each vocabulary to all of the others would be time consuming and inefficient. As a rule, we want to map small vocabularies to the large ones. Small vocabularies are generally used in a specific way by collection specialists while the large vocabularies have more widespread use and contain more synonyms and relations. For this case study we use The Netherlands Institute for Art History (RKD)[5] subject thesaurus as the small source vocabulary. We chose a subject thesaurus because users tend to search on the subject of artworks rather than, say, on materials, therefore, linking it to a vocabulary with more synonyms creates more access points to the collection. For the target thesaurus we chose Cornetto[6] [11], which can be best understood as the Dutch version of Princeton WordNet[7] with additional relations. Both vocabularies are in Dutch, and an extra added value of using Cornetto, in addition to its large coverage, is that it has links to English WordNet.

The original thesauri were in XML format. For project purposes, the RKD thesaurus had already been converted to SKOS[8] and Cornetto to the Princeton W3C schema. The source thesaurus contains 3,576 concepts with 3,342 preferred labels and 475 alternative labels and has broader, narrower and related relations. Cornetto, contains 70,434 synsets and a large number of relations such as hypernym, hyponym and meronym, as well as skos:exactMatch links to the English WordNet. Since the source thesaurus is much smaller than the target thesaurus we are likely to find one-to-many alignments. One benefit of aligning a small thesaurus to a large one, as opposed to aligning large vocabularies to each other, is that, due to the smaller number of possible alignments, the results can be evaluated manually.

Selection of alignment techniques

We selected three alignment techniques and their implementations for generating exact-match relations:

First, a simple syntactic exact match technique to use as a baseline following the strategy used in the OAEI workshop, where a simple edit distance algorithm is used in the benchmark task [5, 2]. We used a homegrown tool for performing exact matching on concept labels as this

[5] http://english.rkd.nl/
[6] http://www2.let.vu.nl/oz/cornetto/index.html
[7] http://wordnet.princeton.edu/
[8] http://www.w3.org/2004/02/skos/

is straightforward to implement. To improve precision the tool ignores all concepts where multiple alignments are possible.

Second, a technique that uses linguistic analysis based on the survey results by Aleksovski et. al.[1]. They propose using lexical techniques for "unified view of collections" type applications: applications with a balanced need for precision and recall and involve knowledge sources of medium complexity such as thesauri. Use cases include access to heterogenous collections, the E-Culture project being a good example. For the implementation of linguistic techniques we used the "in-house" STITCH tool [8] that uses lexical matching techniques such as compound splitting and lemmatization.

Third, a technique that also deploys the ontology's structure. We chose Falcon-AO[9] [7] which uses the structure of vocabularies besides other techniques for finding alignments. It is also "state of the art" giving one of the best performances at the 2007 OAEI workshop and is freely available for deployment on any data-set.

Manual Evaluation
We performed a (time-consuming) manual evaluation of the alignments, that provides a good view of the quality of the alignments. All proposed exact-match relations were rated according as exact-match, incorrect, broader, narrower, related, rejected or "unsure". As the evaluation of all alignments was performed by a single person, we need to rate at least a random sample of the alignments by outsiders in order to get inter-observer agreement statistics by measuring Cohen's Kappa [3]. If the agreement between raters is sufficiently high, the result is a Gold standard, which can be used to evaluate new techniques and to provide guidelines for improving the quality of alignments.

Techniques for improving precision
For large vocabularies, alignments cannot be evaluated manually. We want to develop techniques for improving precision by disambiguating alignments automatically and evaluate the performance of these techniques on the Gold Standard. We aim at reducing the number of one-to-many alignments by removing incorrect alignments using the structure of the vocabularies. An example of an ambiguous alignment is the concept "queen" (royalty) mapped to "queen" (royalty) and "queen" (chess piece). We evaluate two home-brewn techniques for disambiguation described in detail in Section 5.

Study design and data collection
In **Step 1** we preprocess the data-sets by converting them to the formats required by Falcon and the STITCH tool. In **Step 2** we apply the three alignment techniques to our the vocabularies. The tools are used in parallel and independently of each other. We record the time each tool takes to perform the alignments, as well as their ease of use. In **Step 3** we perform manual evaluation of the data by classifying each alignment into one of six categories: exact-match, broader, narrower, related, unsure and rejected. We explain these categories in Section 4. Since this is a time consuming task we record the amount of time the entire process takes. In **Step 4** we have independent raters evaluate a random sample of alignments in order to get inter-rater agreement statistics (Cohen's Kappa). We consider the results of the manual evaluation to be a Gold standard. We can now assess the performance of each tool. We can also assess the added value of combining their results by looking at the amount of overlap between the tools. Here, our focus is on correct exact-match alignments. Finally, in **Step 5** we apply two disambiguation techniques. We measure the number of true positives and false positives filtered out by each of the techniques and also measure the number of false negatives that were removed from the pool of alignments.

[9] http://iws.seu.edu.cn/projects/matching/

4. ALIGNMENT GENERATION

4.1 Preprocessing

Most tools have various preprocessing needs, including the STITCH tool and Falcon-AO. Before using the STITCH tool, Cornetto needed to be converted to SKOS, the RKD thesaurus already being in SKOS format. The wn20s:senselabels were converted to skos:altLabel and hyperonym/hyponym relations to skos:broader/narrower relations. All other relations between synsets were ignored by the STITCH tool.

For Falcon-AO, both vocabularies needed to be converted into an RDF/OWL representation. SKOS labels and senselabels were converted to rdfs:label. As a result, the distinction between preferred and alternative labels was lost in the source thesaurus (RKD). Each concept became an owl:Class and broader/hyperonym relations were converted to rdfs:subClassOf property statements.

4.2 Alignment generation

We generated alignments using the three tools discussed in Section 3. Running the baseline tool took approximately 10 minutes, including loading time of the vocabularies. Alignments were generated using both preferred and alternative labels, with no distinctions being made between the two. To improve precision the tool returned one-to-one alignments only.

Generating alignments with the STITCH tool took approximately 2 minutes. The tool generates one-to-one and one-to-many alignments and aligns nouns and adjectives, not verbs. The tool distinguishes between preferred and alternative labels, alignments based on the latter get a lower confidence rating. Cornetto contains no distinction between labels, while the RKD subject

thesaurus contains both preferred and alternative labels. The alignments were also separated according to the technique used, exact-match with compound splitting and exact-match using lemmatization. The result are four sets of alignments: match on preferred label to alternative label(s) (PrefAlt), match on preferred label lemma to alternative label lemma(s) (PrefAltLemma), match on alternative label to alternative label(s) (AltAlt) and match on alternative label lemma to alternative label lemma(s) (AltAltLemma).

Obtaining results from the Falcon-AO tool took some time. The first runs with varying parameters generated no alignments. Falcon is optimized for English, and Dutch XML language tags in our vocabularies were the reason for finding no alignments. After removing the language tags we ran the Falcon tool with default parameters on the two vocabularies and generated alignments after approximately 20 hours of runtime.

Table 1: Alignments generated

Method	total alignments	source concepts mapped
baseline	1403	1403
PrefAlt	3184	1901
PrefAltLemma	380	176
AltAlt	397	255
AltAltLemma	59	28
STITCH sum	4020	2194
Falcon-AO	2732	2610
Distinct total	4681	2660

Table 1 displays the alignments generated by each tool. The baseline string matching algorithm found the lowest number of alignments, 1403 alignments for 1403 source concepts. This was expected due to the restrictive nature of the technique.

The lexical tool generated 4020 alignments for 2194 source concepts and an average of 2 alignments for each source concept, meaning a large portion of the alignments is ambiguous and possibly incorrect. More than three quarters of the alignments were found using compound splitting and exact matching, with few alignments found using lemmatization. There were also more alignments found for preferred labels than alternative labels of concepts in the source thesaurus due to the higher number of the first type.

Falcon found 2732 alignments for 2610 concepts and has few one-to-many alignments. Falcon aims for higher recall by only returning alignments above a certain treshold. In this respect the STITCH tool is more indiscriminate generating all possible alignments for homonyms.

To investigate the added value of using multiple alignment techniques we need to look at their degrees of overlap. Fig. 1 displays all three techniques in a Venn diagram with the number of alignments in each segment.

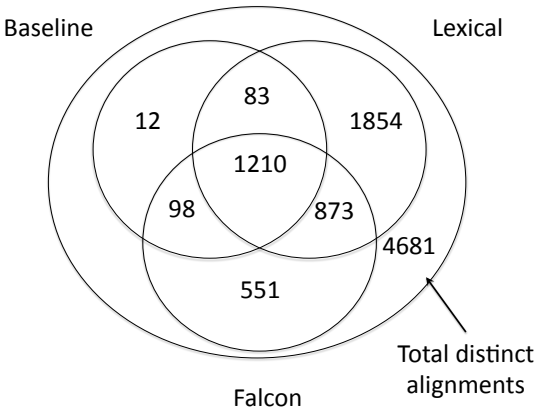

Figure 1: Venn diagram representing alignments per technique and their overlaps

The figure shows that 1210, approximately a quarter of the total of alignments, is found by each of the techniques. These alignments are the easiest to find "low hanging fruit". The segments with no overlap show that the baseline technique adds 12 alignments, less than 1% of the total, on top of what Falcon and the STITCH tool generate. Falcon on the other hand generates 551 extra alignments, 11% of the total, while the STITCH tool 1854 extra alignments, 40% of the total. These numbers seem to confirm the added value of combining techniques for generating alignments.

916 RKD subject thesaurus concepts were not mapped at all to the target thesaurus. A portion of these concepts is formed by multiple words or short sentences and tend to be at the top level of the thesaurus. Examples of these are "levensfasen van de mens" (life-phases of man) and "fysieke en/of psychische toestand (guideterm)" (physical and/or mental state (guideterm)). There were also terms that cannot be found in Cornetto such as "zangvogel" (singing bird) and "scheepsportret" (ship portait). The latter is an example of a domain-specific term found in the source thesaurus, targeting the description of the content of artworks.

4.3 Alignment Evaluation

4.3.1 Manual evaluation of the alignments

In order to assess the quality of the exact-match relations we performed a manual evaluation of each of them. When a concept is mapped to a more specific concept the relation is categorized as narrower, an exact-match to a more general concept is marked as broader. When the concepts are clearly related, such as for example the concept of "Caritas" (the allegory of charity) and "charity", the relation is labeled as related. In some cases the relationship is not clear, often due to ambiguity in the thesauri. In such cases the alignment is categorized

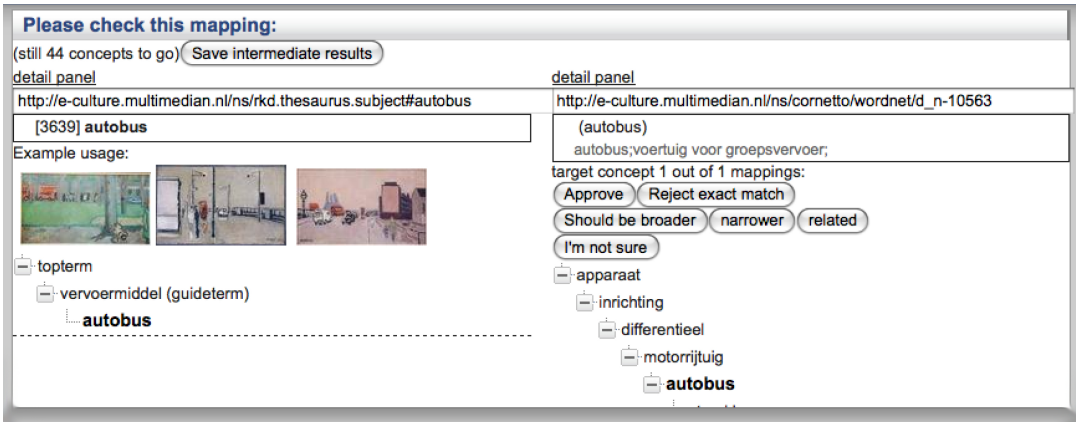

Figure 2: A screenshot of the tool used for evaluating alignments by hand.

as "not sure". Finally, alignments which are evaluated not to be exactMatch nor any of the other relations are marked as rejected.

We created a tool for performing evaluation shown in Fig. 2. For each source concept it displays all the available alignments. The parent concepts are also displayed for each concept, and more information can be viewed about the concept by clicking on "detail panel". If a concept has been used for annotating artworks, up to 5 thumbnails of artworks are displayed to help in the decision process. As the target thesaurus, Cornetto, has not been used for annotation no thumbnails can be displayed for the target concept. For each proposed exact-match relation, the evaluator has to select one of the 6 alignment categories.

Performing the evaluation of 4681 alignments for 2660 RKD subject concepts took slightly longer than 26 man-hours. On average, evaluating a single alignment cost 20 seconds. Correct alignments and obvious rejects took the shortest amount of time, while alignments with other relations generally took a bit longer. In some cases the usage of a source concept was investigated by looking at artworks more closely.

4.3.2 *Validation*

In order to validate the manual evaluation of the alignments, we asked 5 raters to each evaluate alignments for 50 source concepts. The number of alignments varied between 82 and 93 as a single concept can have multiple alignments. Fixing the number of source concepts as opposed to alignments provided a more natural cut-off point as the number alignments per source concept vary. We selected source concepts randomly from the pool of aligned concepts. The raters were provided with guidelines[10]. It took the raters on average 19 minutes to evaluate the alignments. We then compared their ratings with our evaluation of the the same alignments.

[10] `http://e-culture.multimedian.nl/rkd/`

We measured Cohen's Kappa [3] for each external rater for 6 categories. The average of the result kappa's is $\kappa = 0.58$ which is interpreted as moderate agreement. Although the alignments were evaluated over 6 categories, for the purposes of this case study we are mostly interested in correct exact-match relations. Therefore, we also measured Cohen's Kappa for two categories, correct exact-match and an aggregation of the 5 other categories, and measured an average $\kappa = 0.70$ which we find acceptable.

When looking at the disagreements between raters, especially when one rater marks an alignment as "approved" and the other as "rejected", we found two main causes. The first is human error, an alignment was categorized falsely by one or both raters. The second cause is disagreement in the interpretation of the thesauri. In Cornetto, sometimes different meanings of the same concept have not been disambiguated, or concepts are in the wrong hierarchy. While some raters classify alignments as correct even if the meaning is ambiguous, or the concept is in the wrong hierarchy, others reject such alignments. For example the alignment of the concept "spiering", which is a type of fish. In the target thesaurus the fish "spiering" was under the hierarchy for "muscle" but it had a gloss stating it being a sea-fish. In the future we plan to provide guidelines for interpreting such errors in the thesaurus during evaluation.

4.3.3 *Result: Gold standard*

We present statistics per technique on the evaluated and validated alignments in Table 2. The rows "baseline", "STITCH sum" and "Falcon-AO" display the result for each technique that was used. The baseline technique has the highest level of precision of the three techniques with over 93% correct exact-match alignments. The Falcon tool has a precision of 67% and the STITCH tool scores lowest with 53% precision. The results of the STITCH tool are also displayed according to the strategy used. From these details we find that the technique

Table 2: Evaluation results

Method	total alignments	source concepts	exact-match	broader	narrower	related	not sure	rejected
baseline	1403	1403	1309	8	0	46	4	36
PrefAlt	3184	1901	1894	8	23	121	37	1101
PrefAltLemma	380	176	19	2	22	62	2	273
AltAlt	397	255	207	2	21	24	5	138
AltAltLemma	59	28	4	0	2	6	0	47
STITCH sum	4020	2194	2124	12	68	213	44	1559
Falcon-AO	2732	2610	1825	15	20	169	17	686
Distinct total	4681	2660	2384	19	72	290	45	1868

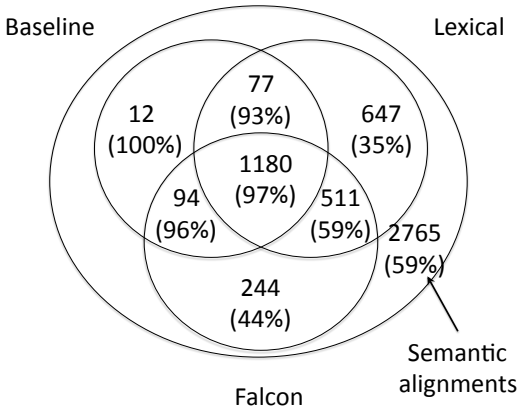

Figure 3: Venn diagram of the validated alignments (exact-match, broader, narrower and related)

using lemmatization has the poorest performance. The average precision over preferred (PrefAltLemma) and alternative labels (AltAltLemma) is 5%, much lower than the precisions of 59% (PrefAlt) and 52% (AltAlt) using exact match with compound splitting. In general alignments found using alternative labels are considered to be less reliable. In this case there is little difference in accuracy which suggests that the two types of label are equally informative in the RKD subject thesaurus.

A small portion of alignments found by the three techniques represent other semantic relationships: broader, narrower and related. For Falcon and the STITCH tool they account for 7% of the total alignments and 4% for the baseline technique.

The last row of Table 2 shows the distinct number of alignments per category. 51% of all alignments is correct exact match and 82% of the source concepts has at least one correct alignment to Cornetto. Another 8% of the alignments represents some semantic link (broader, narrower or related) and 40% of the alignments is not correct.

Fig. 3 displays the number of alignments that overlap between the three techniques categorized as having a semantic relation. That is they are sum of alignments that were evaluated as correct exact-match, broader, narrower and related. The numbers in parentheses display percentages of correctness.

In the baseline technique segment most (94%) of the alignments were exact-match and the other semantic relations only account for less than 4%. The incorrect alignments are homonyms where the correct concept is not part of the target thesaurus. An example is "citroentje" which is a kind of butterfly in the RKD subject thesaurus and an alcoholic drink in Cornetto. The high precision of the baseline technique means that the overlaps with other techniques also have a high precision.

For the alignments found only by Falcon and the STITCH tool 51% of the total is exact-match, and the remaining 8% was categorized as broader, narrower or related. An example of the latter category is the concept "strawberry" meant as the fruit which is related to "strawberry" the plant. Most of the rejected alignments were homonyms as for example "balcony", which is part of a building and "balcony" a type of seating in theatre.

The number of alignments with semantic relations found only by Falcon is 244 which is 44% of the total alignments. Of these, 157 alignments were exact-match and 87 were categorized as broader, narrower or related. In this segment we have a significant percentage of other types of relations besides exact match. We see among exact-match alignments found by Falcon several concepts whose label is composed of multiple terms such as "rode kool" (red cabbage), "boete doen" (paying penance) or compound terms such as "Driekoningenfeest" (Epiphany) matched to "Driekoningen". Similarly, the alignments evaluated as related, broader and narrower are also returned because of partial matches of a word. An example of this a related alignment returned by Falcon only is "geslacht varken" (slaughtered pig) to "varkensslacht" (pigslaughter). The matching of substrings also generates errors. For example "vogelkooi" (birdcage) is matched to "kooivogel" (caged bird) and "streekkleding" (regional clothing/wear) to "strobedekking" (thatch).

The number of alignments with some semantic relation found only by the STITCH tool is higher than for Falcon (647) but represents a smaller percentage (35%) of the total number of alignments. Of these, 475 were exact match, and 172 were broader, narrower or related. Again we see in this segment a significant number of alignments other than exact match.

5. ALIGNMENT DISAMBIGUATION

We can use the Gold standard for evaluating disambiguation techniques aimed at improving precision. The three tools together generated 2966 alignments for 955 concepts which is an average of 3 alignments per concept. An example of ambiguous alignments is the concept "king" as royalty in RKD thesaurus which has three alignments. The first alignment is to a playing card "king", the second is to the chess piece "king" and the third is to royalty "king". The first two alignments are false positives and we need to some disambiguation technique to detect them. One option could be to check whether the parent concepts match. This is only the case for the third alignment where the broader term for both source and target concept is "vorst" which means ruler. We have implemented two disambiguation techniques exploiting the structure of thesauri: disambiguation by counting "child" alignments (Child Match), and by counting "parent" alignments (Parent Match).

5.1 Disambiguation Techniques

The goal of these techniques is to reduce ambiguity by establishing correct exact-match alignments based on the amount of alignments in the lower or upper levels of the hierarchies.

In the Child Match technique, for each concept with multiple alignments we follow the hierarchy "down" using narrower relations and count the number of alignments in the lower reaches between the two vocabularies. We assume that concepts which are equal in meaning will have similar hierarchies below them. This means there are more alignments between their children, than for concepts which may be lexically similar but differ in meaning. We then count the number of alignments that have at least one or more child alignments and consider them to be correct exact-match. If multiple alignments for a single concept have more than one child alignment we choose the alignment with the highest number of child alignments. However, in some cases both alignments have the same (highest) number of child alignments and then both are chosen.

Parent Match is a mirroring of the Child Match counting technique. We want to find correct alignments by exploiting the top of the hierarchies. For each ambiguous alignment we count the number of alignments that could be reached from each concept through broader relations. Alignments with at least one "parent" alignments are considered to be correct exact-match.

5.2 Results

Table 3: Results of disambiguation techniques applied to 955 concepts with 2966 alignments

	Disamb. concepts	Alignm. kept	Alignm. removed	true pos.	false pos.	false neg.
Child Match	115	123	331	93	30	33
Parent Match	183	231	322	181	50	42
Distinct total	280	336	605	247	79	74

Table 3 displays the results of the implementation of both techniques. Using the Child Match technique we found 125 alignments for 115 source concepts with at least one child alignment. Our assumption is that these 125 alignments are correct and therefore we removed 331 alignments that had fewer or no child alignments for the same 115 source concepts. We evaluated the effect of this technique using the Gold standard. We found that 93 of the 125 alignments were correctly selected (true positives). We also counted the number of false negatives, or alignments that were removed but that are in fact correct. We found that a little over 10%, that is 33 out of 331 removed alignments were false negatives. Examining the false negatives we found that the main reason for excluding them was because they had no child concepts or the child concepts are organized differently. For example the concept "factory" in the RKD thesaurus has multiple child concepts such as "steel factory" and "brickyard" while in Cornetto the child concept is "factory hall".

Using the Parent Match technique we found 231 alignments for 183 source concepts. We removed 322 alignments that had no parent matches for these 183 concepts. Again we used the Gold standard to evaluate the results of this technique and found that 181 of the 231 alignments were true positives. We also examined the alignments removed by this technique and found that 42 out of 322 alignments were false negatives. The reason these alignments are not returned is usually because of differences in hierarchies. For example the concept "almanac" has as parents "book" and "printed work" in the RKD thesaurus and "expression" "description" and "chronicle" in Cornetto. There is a small overlap in the alignments found by the Child and Parent Match techniques. Overall the two disambiguation techniques together reduced the number of ambiguous alignments by a third.

6. DISCUSSION

In this case study we have applied and evaluated a number of typical state-of-the-art techniques for ontology alignment. Now, can we draw useful methodological lessons from this case study?

The three alignment-generation techniques found align-

ments for 75% (2,660) of the 3,576 source concepts. We have not studied in detail the remaining 25%, but manual inspection of a random sample of 30 concepts showed that 26 of these had a direct semantic link (broader, narrower or related) with a concept in the 75% set. For the application context of this study (the alignments are used as part of a semantic network for information retrieval, the E-Culture semantic search engine) such an outcome is fine and would not warrant spending a lot of time on the rest-group.

Table 4: Precision, recall and F-measure values

segment	align-ments	exact-match align.	precision	recall	F-measure
1: baseline	1403	1309	0.93	0.55	0.69
2: 1+ overlap	2256	1739	0.77	0.73	0.75
3: 1+ disamb. overlap	2111	1714	0.81	0.72	0.76
3 + missing concepts	2896	1909	0.66	0.8	0.72
3 + manual evaluation of missing concepts	2317	1909	0.82	0.8	0.81

A second observation is that, when only considering exact-match alignments, the baseline method has the highest precision: 1309 correct alignments are found for 1403 distinct concepts, giving 93% precision, at the cost of a lower recall of 55% (F-measure=0.69) shown in Table 4. If an alignment is only found by one of the two other techniques, precision drops to well below 30%; alignments included in the intersection of the two methods have a reasonable precision (50%). This result improves to 57% when the disambiguation techniques are applied. Combining the disambiguated overlap between the two other methods and the baseline results in a lower precision of 81%, but a significantly higher recall at 72% (F-measure=0.76).

From these findings we can hypothesize that the following alignment procedure might be the optimal:

1. Apply the baseline method and accept all results.
2. Apply a lexical and a structured technique to find overlapping results.
3. Apply the disambiguation techniques to this overlap and accept the results.

For those applications where 72% recall is insufficient, the recall can be improved by adding alignments found only by falcon and the lexical techniques. To prevent a drop in precision these alignments should be evaluated manually. We then only need to evaluate alignments for concepts for which no alignments were found by the baseline method and the disambiguated overlap. In this case there would be 785 alignments, which, according to our experiences, would take approximately 4 person-hours to evaluate. This investment would boost recall to 80% and slightly improve precision to 82% (F-measure=0.81). Note that here we have been very strict, and have counted proposed exact-match relations that were evaluated as broader, narrower or related as incorrect as if they had no relation at all.

7. ACKNOWLEDGEMENTS

The datasets have been kindly provided by RKD and the Cornetto project. We thank Antoine Isaac and the STITCH project, Borys Omelayenko and Wei Hu for helping us using their tools, and Mark van Assem, Willem van Hage, Laura Hollink and Jan Wielemaker for their contributions on the alignment evaluation. This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government.

8. REFERENCES

[1] Z. Aleksovski, W. V. Hage, and A. Isaac. A survey and categorization of ontology-matching cases. In P. Shvaiko, J. Euzenat, F. Giunchiglia, and B. He, editors, *Proceedings of the Workshop on Ontology Matching (OM2007) at ISWC/ASWC2007, Busan, South Korea*, November 2007.

[2] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, and V. Svátek. Results of the ontology alignment evaluation initiative 2008. In *OM*, 2008.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960.

[4] J. Euzenat, M. Ehrig, A. Jentzsch, M. Mochol, and P. Shvaiko. Case-based recommendation of matching tools and techniques. deliverable 1.2.2.2.1, Knowledge Web NoE, 2006.

[5] J. Euzenat, A. Isaac, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, O. Sváb, V. Svátek, W. R. van Hage, and M. Yatskevich. Results of the ontology alignment evaluation initiative 2007. In *OM*, 2007.

[6] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[7] W. Hu and Y. Qu. Falcon-ao: A practical ontology matching system. *Web Semant.*, 6(3):237–239, 2008.

[8] A. Isaac, C. T. dos Santos, S. Wang, and P. Quaresma. Using quantitative aspects of alignment generation for argumentation on mappings. In P. Shvaiko, J. Euzenat, F. Giunchiglia, and H. Stuckenschmidt, editors, *OM*, volume 431 of *CEUR Workshop Proceedings*, 2008.

[9] M. Nagy, M. Vargas-Vera, P. Stolarski, and E. Motta. Dssim results for oaei 2008. In *OM*, 2008.

[10] G. Schreiber, A. Amin, L. Aroyo, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, B. Omelayenko, J. van Osenbruggen, A. Tordai, J. Wielemaker, and B. Wielinga. Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator. *Web Semant.*, 6(4):243–249, 2008.

[11] P. Vossen, I. Maks, R. Segers, and H. VanderVliet. Integrating lexical units, synsets and ontology in the cornetto database. In E. L. R. A. (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[12] X. Zhang, Q. Zhong, J. Li, and J. Tang. Rimom results for oaei 2008. In P. Shvaiko, J. Euzenat, F. Giunchiglia, and H. Stuckenschmidt, editors, *OM*, volume 431 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.