Applied Ontology 0 (2008) 1–0
IOS Press

The OAEI food task: an analysis of a thesaurus alignment task

Willem Robert van Hage [a,b,*] Margherita Sini [c] Lori Finch [d] Hap Kolb [a] Guus Schreiber [b]

[a] *TNO Science & Industry, Stieltjesweg 1, 2628 CK, Delft, the Netherlands*
[b] *Vrije Universiteit, de Boelelaan 1081a, 1081 HV, Amsterdam, the Netherlands*
[c] *Knowledge Exchange and Capacity Building Division, Food and Agriculture Organization of the United Nations (FAO), Viale delle Terme di Caracalla, 00153 Rome, Italy*
[d] *National Agricultural Library, United States Department of Agriculture (USDA), 10301 Baltimore Blvd., Beltsville, MD, USA*

Abstract. This paper describes the "food task" of the Ontology Alignment Evaluation Initiative (OAEI) 2006 and 2007. The OAEI[**] is a comparative evaluation effort to measure the quality of automatic ontology-alignment systems. The food task focuses on the alignment of thesauri in the agricultural domain. It aims at providing a realistic task for ontology-alignment systems by which the relative performance of the alignment systems can be evaluated. Research groups from around the world signed up their ontology-alignment system for the task. Each system automatically constructed an alignment. The alignments were then compared by means of statistical performance measures to get clues about which techniques work best for automatic ontology alignment. To complement this quantitative evaluation we performed an in-depth qualitative analysis of the results to draw conclusions about the strengths and weaknesses of the various alignment approaches and the specific challenges of thesaurus alignment and its evaluation.

1. Introduction

Ontology alignment has become a major research focus in the area of distributed Web applications. The Web has has made it possible to access multiple libraries at the same time. Different libraries have different indexing schema's. This makes federated access difficult. In the past, this was solved by unifying the schema's. This can fail when there are non-reconcilable differences between the schema's or conflicts of interest. Alignment can be seen as an alternative to schema unification, *cf.* (Clarke, 1996). The schema's stay unchanged; instead cross-links between the the schema's are added. Differences between the schema's are allowed to persist. (Huang et al., 2005, 2006) The alignment has to be maintained, but this is a smaller issue to solve than to arrange joint maintenance of a unified schema.

Initially, OAEI focused on alignment of heavy-weight OWL-based ontologies. However, in practice the domains in which alignment is needed are typically information retrieval tasks where documents (including multimedia documents such as images and video) have been indexed with different thesauri. Such concept schemes can best be viewed as light-weight ontologies. Many thesauri follow the ANSI/NISO and ISO standards for thesauri, such as ANSI/NISO Z39.19, ISO 2788 (for monolingual thesauri) and ISO 5964 (for multi-lingual thesauri) Hodge (2000). Within the Semantic Web community SKOS (Simple Knowledge Organization System) has been developed for the purpose of providing a format for publishing such thesauri on the Web. SKOS Miles and Bechhofer (2004) allows one to define a concept scheme with a URI for each concept so that we can create unambiguous alignments between the thesauri. SKOS provides a special alignment vocabulary, the SKOS Mapping Vocabulary (discussed in more detail in Section 2.3).

The main research objective of this paper concerns the *evaluation methodology for ontology alignments*. We use the results of the OAEI 2006/2007–a comparative evaluation challenge for ontology matching

*Corresponding author: Tel.: +31-20-5987751; E-mail: wrvhage@few.vu.nl.
**http://oaei.ontologymatching.org/

systems, where participants are allowed to use any algorithm they can implement, to align given ontologies and try to outperform the other participants–as a case study to get insight into evaluation issues such as the way in which recall and precision should be assessed. In real-life alignment cases (of which the food task is an example) there is often no gold standard for the alignment available. We are also interested in characteristics of thesaurus alignment in comparison with general ontology alignment.

We start by explaining the data involved in the OAEI 2006 and 2007 food task. Section 2 describes the vocabularies that were used and Section 3 describes the alignments submitted by the participants. Subsequently, we discuss in Section 4 the general evaluation method that we followed. In Sections 4.1 and 4.2 we elaborate on the specific details of the OAEI 2006 and 2007 food task evaluation. In Section 5 we quantitatively compare the performance of the participating systems. Finally, in Section 6 we perform a qualitative analysis of the results, where we discuss in some detail typical issues with respect to alignment of thesauri.

2. Vocabularies

The thesauri used for this task are the United Nations Food and Agriculture Organization AGROVOC thesaurus, and the United States National Agricultural Library Agricultural Thesaurus. We selected these thesauri because they are both large and widely used. The thesauri were supplied unaltered to the participants in their native SKOS format and a simplified OWL-Lite version. The 2006 OWL-Lite version was made by Wei Hu. The 2007 OWL-Lite version follows the same rules as those used by Antoine Isaac for the OAEI 2007 library track.[1] The versions used for the OAEI 2006 and 2007 food task can be downloaded at `http://www.few.vu.nl/~wrvhage/oaei2006` and `http://www.few.vu.nl/~wrvhage/oaei2007/food.html` respectively.

2.1. AGROVOC

The UN FAO AGROVOC thesaurus was developed by agriculture domain experts at the FAO and the Commission of the European Communities, in the early 1980s. It is updated by FAO roughly every three months. AGROVOC is used to index a multitude of data sources all over the world, one of which is the AGRIS/CARIS[2] literature reference database. Many international organizations use localized excerpts of the thesaurus. Information about these projects and links to the respective web pages can be found at `http://www.fao.org/aims`. There are manually created alignments from AGROVOC to the Chinese Agricultural Thesaurus and the German National Library's Schlagwortnormdatei, and an automatically generated alignment to the European Environment Agency's GEMET.[3] AGROVOC is available in many different formats including ISO 2709 (format for bibliographic information interchange), SKOS, OWL,[4] and TermBase eXchange (TBX).[5] All formats are generated from a native custom MySQL form. The current version of AGROVOC thesaurus can be browsed online at `http://www.fao.org/agrovoc`. An online collaborative maintenance system for AGROVOC, called the AGROVOC Concept Server Workbench, is under development.[6] Future versions of the thesaurus will also be made available through a web service.

For the OAEI 2006 food task we used the May 2006 version which consists of 28,174 descriptor terms (*i.e.* prefered terms) and 10,028 non-descriptor terms (*i.e.* alternative terms). It is multilingual in ten languages (English, French, Spanish, Arabic, Chinese, Portugese, Czech, Japanese, Thai, and Slovak). For the OAEI 2007 food task we used the February 2007 version which consists of 28,445 descriptor terms and

[1] `http://www.few.vu.nl/~aisaac/oaei2007`
[2] `http://www.fao.org/agris`
[3] `http://www.few.vu.nl/~wrvhage/oaei2007/environment.html`
[4] `http://www.w3.org/2004/OWL`
[5] `http://www.lisa.org/standards/tbx`
[6] `http://www.fao.org/aims/aos.jsp`

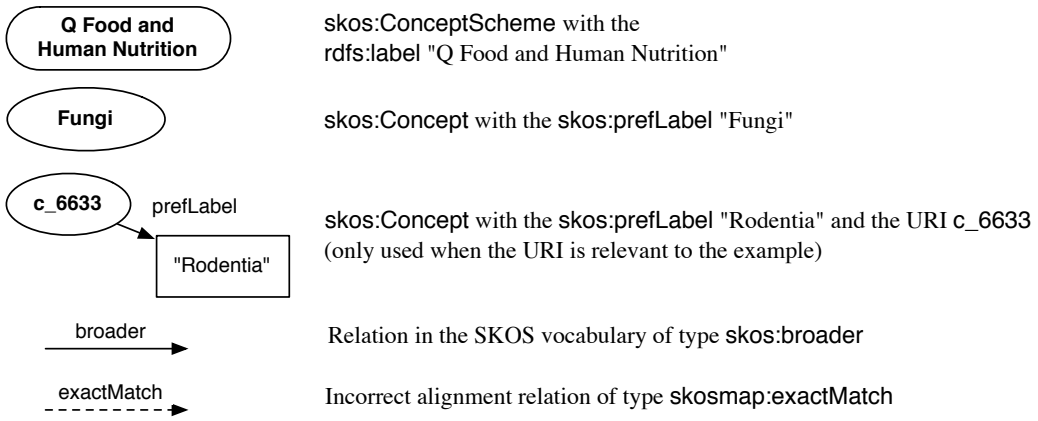

Fig. 1. Legend to the visual symbols used in this paper.

12,531 non-descriptor terms and is multilingual in eleven languages (the same as listed before, plus German). Strictly speaking, AGROVOC is a translated thesaurus and not a multilingual thesaurus. It started with an English version and was later translated into other languages by domain experts from the respective countries. The terms are grouped into categories from the AGRIS/CARIS Classification Scheme.[7]

The SKOS format has exactly one skos:Concept per descriptor term. The term itself is a skos:prefLabel of the skos:Concept. Non-descriptors (USE) are modelled as skos:altLabels. USE+ is downgraded to multiple unrelated skos:altLabel relations. BT, NT, and RT relations are modelled as skos:broader, skos:narrower, and skos:related relations between the respective skos:Concepts. AGRIS/CARIS Classification Scheme categories are modelled as skos:ConceptSchemes. The broadest concept that has a AGRIS/CARIS classification is modelled as a top concept of that skos:ConceptScheme using skos:hasTopConcept. Whenever scope notes exist they are attached to the skos:Concept as strings using the skos:scopeNote property.

An excerpt of AGROVOC is shown in figure 2 on the left side. In all figures in this paper we will depict skos:Concepts as an oval filled with the skos:prefLabel text. In cases where we explicitly want to show skos:altLabel and skos:prefLabel we depict the skos:Concept as an oval filled with its URI, connected to boxes that represent its various labels. skos:ConceptShemes are depicted as boxes with round sides. An overview of these visual symbols is shown in Figure 1.

2.2. NAL Agricultural Thesaurus

The USDA NAL Agricultural Thesaurus (NALT) was created by the National Agricultural Library to disclose information of the Agricultural Research Service of the United States Department of Agriculture. In 2002 the first English edition was published. In 2007 the first Spanish version of the NALT was published. Both are updated annually. The NALT is used to index the AGRICOLA[8] literature reference database of the USDA, the Food Safety Research Information Office[9] (FSRIO) research projects database, the NAL Digital Repository[10] (NALDR), and various data sources of the Agriculture Network Information Center[11] (AgNIC). There is an automatically generated alignment to the European Environment Agency's GEMET thesaurus. NALT is available in SKOS, and MARC, and a custom ASCII and XML format. The SKOS format is generated from the XML format. This transformation follows the same rules as described above for the SKOS version of AGROVOC. The current English version of the NALT thesaurus can be browsed online at `http://agclass.nal.usda.gov/agt/agt.shtml`. More information about

[7] `http://www.fao.org/aims/ag_classifschemes.jsp`
[8] `http://agricola.nal.usda.gov`
[9] `http://fsrio.nal.usda.gov`
[10] `http://naldr.nal.usda.gov`
[11] `http://www.agnic.org`

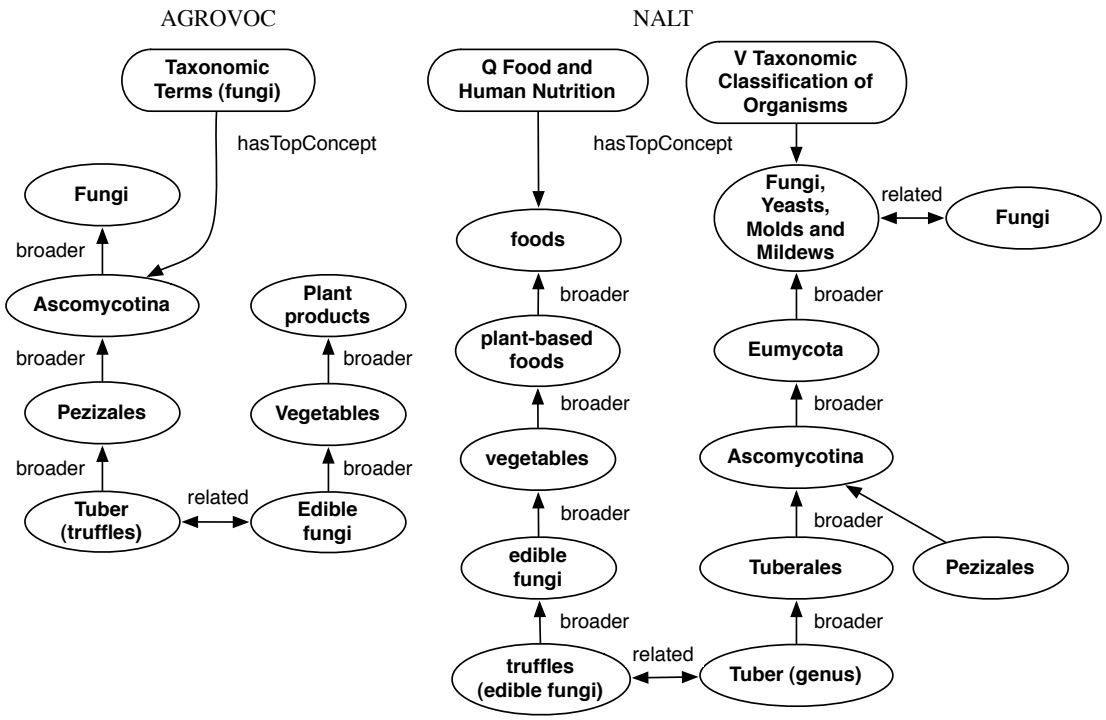

Fig. 2. The concept representing truffles in AGROVOC and NALT.

the Spanish version can be found online at `http://agclass.nal.usda.gov/agt_Espanol/agt_es.shtml`.

For the OAEI 2006 food task we used the 2006 version of the NALT which consists of 41,577 descriptor terms and 24,525 non-descriptor terms and is English monolingual. For the OAEI 2007 food task we used the 2007 version, which consists of 42,326 descriptor terms and 25,985 non-descriptor terms. We only use the English version.

An excerpt of NALT is shown in figure 2 on the right side.

2.3. SKOS Mapping Vocabulary

For the alignment we use relations from the SKOS Mapping Vocabulary. The SKOS Mapping Vocabulary specification can be found at `http://www.w3.org/2004/02/skos/mapping/spec`. The participants were allowed to use the following relations: skosmap:narrowMatch, skosmap:exactMatch, and skosmap:broadMatch. The other relations and boolean combinators (skosmap:minorMatch, skosmap:majorMatch, skosmap:AND, skosmap:OR, skosmap:NOT) of the SKOS Mapping Vocabulary were not used in the evaluation. The participants were requested to hand in an RDF file in alignment format[12] (Euzenat, 2004) that contains information about the properties of the alignment, like which ontologies are involved, and properties of each relation in the alignment, like which concepts are aligned and the confidence the participant's ontology alignment system gave to the relation. An example of such an RDF file is shown in the code listing in figure 3. The example shows two alignment relations, nalt:osteomyeliti skosmap:exactMatch agrovoc:c_12988 (Osteomyelitis), and favism skosmap:exactMatch agrovoc:c_6051 (Poisoning). The relations get a confidence of respectively 1.0 and 0.89.

[12] `http://alignapi.gforge.inria.fr`

```
<?xml version='1.0' encoding='utf-8'?>
<rdf:RDF xmlns='http://knowledgeweb.semanticweb.org/heterogeneity/alignment'
         xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
         xmlns:xsd='http://www.w3.org/2001/XMLSchema#'>

  <Alignment>
    <xml>yes</xml>
    <level>0</level>
    <type>11</type>
    <onto1>http://agclass.nal.usda.gov/nalt/2007.xml</uri1>
    <onto2>http://www.fao.org/aos/agrovoc</uri2>
    <map>
      <Cell>
        <entity1 rdf:resource='http://agclass.nal.usda.gov/nalt/2007.xml#
            osteomyelitis'/>
        <entity2 rdf:resource='http://www.fao.org/aos/agrovoc#c_12988'/>
        <measure rdf:datatype='http://www.w3.org/2001/XMLSchema#float'>1.0</measure>
        <relation>http://www.w3.org/2004/02/skos/mapping#exactMatch</relation>
      </Cell>
    </map>
    <map>
      <Cell>
        <entity1 rdf:resource='http://agclass.nal.usda.gov/nalt/2007.xml#favism'/>
        <entity2 rdf:resource='http://www.fao.org/aos/agrovoc#c_6051'/>
        <measure rdf:datatype='http://www.w3.org/2001/XMLSchema#float'>0.89</measure>
        <relation>http://www.w3.org/2004/02/skos/mapping#exactMatch</relation>
      </Cell>
    </map>
    ...
  </Alignment>
</rdf:RDF>
```

Fig. 3. The RDF format used for the submission of alignments. This example shows two skosmap:exactMatch relations with a confidence of respectively 1.0 and 0.89.

3. Participants and Submitted Alignments

The OAEI 2006 food task had five participants: South East University with the Falcon-AO 0.6 system (Hu et al., 2006); University of Pittsburgh with the Prior system (Mao and Peng, 2006); Tsinghua University with the RiMOM system (Li et al., 2006); University of Leipzig with the COMA++ system (Massmann et al., 2006); and Universitá degli Studi di Milano with the HMatch system (Castano et al., 2006). Each team provided between 10,000 and 20,000 alignment relations. This amounted to 31,112 unique alignment relations in total. All of these mappings were of the type skosmap:exactMatch. None of the systems was able to discover skosmap:broadMatch or skosmap:narrowMatch mappings. There was a high agreement between the best three systems, RiMOM, Falcon-AO, and HMatch. Details are shown in table 1. From this table we can also deduce that there is a relatively large set of "easy" mappings that are recognized by all systems.

The OAEI 2007 food task also had five participants: South East University with the Falcon-AO 0.7 system (Hu et al., 2007); Tsinghua University with the RiMOM system (Li et al., 2007); Politecnico di Milano with the X-SOM system (Curino et al., 2007); and the Knowledge Media Institute with two systems, DSSim (Nagy et al., 2007) and SCARLET (Sabou et al., 2007). Each team provided between 6583 (X-SOM) and 18,420 (RiMOM) alignment relations. This amounted to 37,384 unique alignment relations in total. The SCARLET system discovered skosmap:exactMatch, skosmap:broadMatch, and skosmap:narrowMatch relations. The other systems only discovered skosmap:exactMatch relations. There was a slightly lower agreement between RiMOM and Falcon-AO (the Jaccard similarity coefficient, $|A \cap B| / |A \cup B|$, was $11,203/22,517 = 0.50$ as opposed to $11,585/26,984 = 0.75$ in 2006). The other systems found much more different sets of alignment relations than the other systems in 2006. The SCARLET system is a complete outlier compared to the other systems.

2006

system	# mappings returned	# mappings shared with *n* other systems 0	1	2	3	4
RiMOM	13,975	868	1,042	2,121	4,389	5,555
Falcon-AO	13,009	642	419	1,939	4,400	5,555
Prior	11,511	1,543	1,106	676	2,631	5,555
COMA++	15,496	11,610	1,636	629	2,028	5,555
HMatch	20,001	7,000	981	2,045	4,420	5,555
all systems	31,112	21,663	2,592	2,470	4,467	5,555

2007

system	# mappings returned	# mappings shared with *n* other systems 0	1	2	3	4
RiMOM	18,419	7,052	6,131	3,774	1,462	0
Falcon-AO	15,300	2,964	6,933	3,941	1,462	0
X-SOM	6,583	4,083	317	725	1,458	0
DSSim	14,962	9,273	876	3,353	1,460	0
SCARLET exactMatch	81	9	27	39	6	0
broadMatch & narrowMatch	6,038	6,038	0	0	0	0
all systems	41,967	29,419	7,142	3,944	1,462	0

Table 1

Distribution of the systems' results. Shown are the number of mappings returned by each system and how many mappings are also returned by *n* of the other systems.

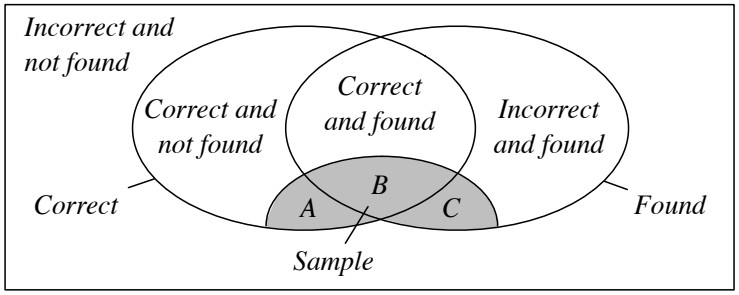

Fig. 4. Venn diagram to illustrate the sets of relations that are relevant to the sample evaluation. $A \cup B$ is a sample of the population of Correct alignment relations. $B \cup C$ is a sample of the population of Found alignment relations.

4. Evaluation Procedure

In this section we will describe the evaluation process we used to compare the various submissions. The main two statistics we used to compare the alignments are Precision and Recall. If we call the set of all alignment relations that were submitted by a participant *Found* and the set of all alignment relations we would like to receive (*i.e.* all correct alignment relations) *Correct*, Precision and Recall can be defined as follows:

$$\text{Precision} = \frac{|Found \cap Correct|}{|Found|} \tag{1}$$

$$\text{Recall} = \frac{|Found \cap Correct|}{|Correct|} \tag{2}$$

Figure 4 illustrates these definitions. In practice, the computation of Precision and Recall require the assessment of all relations in the set of *Found* relations and the determination of the cardinality of the set of all *Correct* relations.

The assessment of all *Found* relations requires human assessors to decide whether tens of thousands of alignment relations are correct or incorrect. The experience of the OAEI has shown that a voluntary

human assessor can judge around 250 alignment relations per hour for at most a few hours. That means 10,000 alignments cost around 40 man-hours. For most large organizations that want to know the quality of an ontology alignment system this is a feasible investment. For evaluation fora such as the OAEI, this is not feasible. For the comparative evaluation of multiple systems we even have to assess multiple sets of *Found* relations.

The assessment of all *Correct* requires the manual construction of the entire desired alignment. Manual construction of the entire alignment is even more costly than the assessment of all *Found* relations, because it involves searching for good alignment relations, which is more difficult than simply judging the validity of a set of given relations. To illustrate this we can look at the manual construction of the alignment between the Chinese Agricultural Thesaurus (CAT), which consists of 64,638 concepts, and AGROVOC. This alignment is directional from CAT to AGROVOC and hence not complete, and consists of 24,686 alignment relations. Chang Chun of the Chinese Academy of Agricultural Sciences (CAAS) revealed at the Eighth Agricultural Ontology Service (AOS) meeting[13] that the construction took 15 PhD students (in relevant fields of research, like biology) 24 man-hours each during 6 months. The students were paid per alignment and followed a strict protocol. They made at most around 150 alignment relations per hour. (Liang et al., 2005)

If you are not interested in the evaluation as such, but in a complete alignment, automatic ontology alignment might not be necessary, because the total investment for the manual construction of an alignment is, for many purposes, not significantly larger than that of verifying an automatically constructed alignment. Provided that time, money, and access to adequately educated people are not an issue. In these cases manual ontology alignment might be worth the investment.

To make the computation of Precision and Recall feasible for the OAEI food task, we performed sample evaluation. Sample evaluation assumes that measurements on a randomly drawn sample can be extrapolated to the entire population. The larger the sample, the less the estimation based on the sample will deviate from the true value on the entire population. In our case, that means that we can extrapolate the performance of a system on a small set of alignment relations to all relevant alignments. We work with small subsets of all *Found* and *Correct* relations from which we generalize to the entire set of *Found* or *Correct* relations. The grey areas $B \cap C$ and $A \cap B$ in Figure 4 illustrate the samples used for the evaluation of respectively Precision and Recall. In Section 4.1 and 4.2 we will go into detail on how these samples were constructed and how the human judges operated exactly.

Sample evaluation comes with a price. It introduces sampling error, bias due to the accidental inclusion and omission of certain elements from the population in the sample. The smaller the sample is, the more likely it is that important features of the population are accidentally overlooked. For instance, we know that the automatic alignment of concepts that represent the animal species is quite simple compared to the alignment of concepts that represent socio-economic phenomena. If a random sample of alignments by accident overrepresents animal species then the performance estimate based on this sample will be too optimistic. The fact that there are many potential alignment relations between animal species and few between socio-economic phenomenae even makes it quite likely that a random sample from all alignment relations contains no socio-economic relations, but quite a few animal species relations. To minimize this kind of bias, we did a separate evaluation for sets of alignment relations that we know in advance to require different alignment strategies. The separate results are combined into a weighted average to give a fair overall performance indication. The statistical technique we used to accomplish this for Precision and Recall were different. For Precision we use stratified sampling, while for Recall sampling we use cluster sampling. (Cochran, 1977) The main reason for this difference is that the set of all *Found* alignment relations, as opposed to all *Correct* alignments, is predetermined. Hence we can easily draw samples from it.

In order to draw samples from the set of all *Correct* alignments we have to draw from the set of *all* alignment relations and filter out the incorrect alignments. Clearly, some parts of the cartesian product of the sets of terms from the two thesauri will contain more correct alignment relations than others (*e.g.*

[13] `http://www.fao.org/aims/pub_aos8.jsp`

there are bound to be matches between the parts about plants of both thesauri, but not between the part about plants of one thesaurus and the area about countries of the other). So if we want to use our time optimally–which we have to do to make the evaluation feasible–we will look for correct alignment relations in the areas that are likely to contain some and not in the areas that are unlikely to contain any. This concession breaks one of the assumptions of stratified sampling, the assumption that the *entire* population is partitioned and that all elements get an equal chance to be selected for a sample.

The closest thing to stratified sampling that does not make assumptions we can not meet is cluster sampling where the clusters are not selected randomly. The price we pay for the reduction in assessment time is that we have no indication how large the error margin is when we generalize from the samples to the entire population of *Correct* alignment relations.

4.1. Precision

We estimate Precision using stratified sampling from the set of all *Found* alignment relations. This set is different for the 2006 and 2007 food task and different for each participating system. We discounted the effect of two kinds of features in the evaluation: how many systems submitted a certain relation, and the topic of the relation. The intuition behind this is the following. It can be expected that the quality of alignment relations that are submitted by all systems and relations that are submitted by, for instance, only one system will be different. It can also be expected that some topics are easier than others, as we explained in the beginning of Section 4 about terms representing animal species.

We first partitioned the set of all *Found* alignment relations into strata with a different topic. All relations between concepts that fall into these topics were grouped together. In 2006 we distinguished three categories of topics in the thesauri that each required a different level of domain knowledge of the assessors: *taxonomical* concepts (plants, animals, bacteria, etc.) that can be aligned with a few simple rules and lexical matching, *biological and chemical* concepts (structure formulas, terms from generics, etc.) that contain many synonyms and lexical variants, and *miscellaneous*, the remaining concepts (geography, agricultural processes, etc.) that can be expected to require a diverse set of techniques to match. In 2007 we distinguished four categories op topics. The same as used in 2006 plus *geographical* concepts (countries, provinces, etc.). We chose to separate these from the *miscellaneous* set, because there is much consensus about the naming of geographical locations. This makes the alignment of geographical concepts much easier than other topics in the *miscellaneous* set.

From each of the sets shown in Table 1 we took a random sample from each of the topic strata, such that both commonly and rarely returned alignment relations would be represented in each topic. Together, this led to the samples shown in Table 2, that had to be assessed.

Under the authority of taxonomists at the USDA the taxonomical stratum was automatically assessed completely using the strict rules that apply to the naming scheme of taxonomy. These rules are that if the preferred term of concept *A* is literally the same as either the preferred or the alternative term of concept *B* then the concepts are considered to be equivalent, provided that the same goes for an ancestor of *A* and *B*. This is illustrated in Figure 5. This assumes that the same taxonomical names always signify the same species, group, kingdom, or the like. In general, this is not true for taxonomical names, but only for names that are disambiguated by the last name of the author that first published the classification and year of the publication, *e.g.* "Passer domesticus (Linnaeus, 1758)". An example of homonymy in species names is "Cereus", which can be either a cactus or sea anemone. In the case of NALT and AGROVOC, however, this ambiguation is not necessary, because the species names were based on the same literature and many of the concepts were copied from the same sources. Therefore, if the terms match it is extremely likely that they refer to the same species.

The other strata were all manually assessed by a group of domain experts. In 2006 this was done by domain experts of the NAL and the FAO, and a group of computer scientists at the EKAW workshop. In 2007 it was done by domain experts of the NAL, FAO, TNO Quality of life, Unilever, Wageningen Agricultural University, and the European Environment Agency. The assessed samples can be downloaded

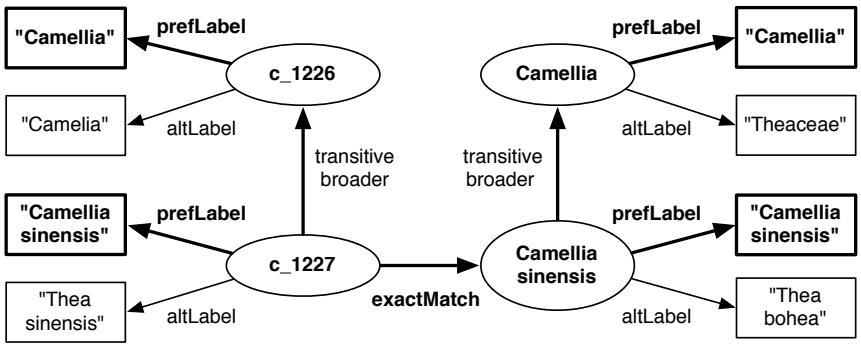

Fig. 5. Automatic assessment of taxonomical terms. The two concepts representing Camellia sinensis are considered equivalent, because they have a matching label and some of their ancestors also have matching labels.

2006

stratum topic	stratum size (N_h)	sample size (n_h)	stratum weight
taxonomical	18,399	18,399	0.59
bio/chem	2,403	250	0.08
miscellaneous	10,310	650	0.33
all topics	31,112		

2007

stratum topic	stratum size (N_h)	sample size (n_h)	stratum weight
taxonomical	23,023	23,023	0.55
bio/chem	3,965	200	0.09
geographical	1,354	86	0.03
miscellaneous	13,625	476	0.32
all topics	41,967		

Table 2

Sizes of the strata and of the samples from those strata that were assessed to evaluate Precision. The last column shows how much the stratum weighed in the calculation of a system's mean Precision.

from `http://www.few.vu.nl/~wrvhage/oaei2006/gold_standard` and `http://www.few.vu.nl/~wrvhage/oaei2007/gold_standard`.

Assessment Tool for Precision For the assessment we used an alignment assessment tool developed at TNO by Willem Robert van Hage. An adaptation of this tool was also used for the assessment of the OAEI 2007 library task. A screengrab is shown in figure 6. This tool reads a set of mappings in the common format for alignments and outputs a web form that is used by judges to assess the mappings. The results of the form are submitted to the organizer of the food task. The assessment process of a mapping follows three steps.

1. The judge decides, based on the prefered label, alternative labels, and the broader terms of the concept, if the relation specified above the arrow (between the two green boxes) holds between the two bold concepts. If the relation holds he skips to point 2 and goes straight to point 3. If it does not hold he goes to point 2;
2. The judge tries to specify an alternative relation, either by changing the relation type, or the concepts. If possible he select "exactMatch" and specifies the proper concepts between which the "exactMatch" relation holds. Otherwise he selects "broadMatch" or "narrowMatch" and specifies the proper concepts between which that relation holds.
3. The judge changes the default value of the assessment, "unknown", into either "true" or "false". If the relation holds and he arrived here from point 1 he selects "true". If the relation does not hold, but if he successfully selected an alternative relation (at point 2) that does hold, he also select "true". If the relation does not hold and no correct alternative could be found at point 2, select "false".

Finally, if the judge wishes to document his decision he fills out the entry box at the bottom of the assessment form. This description was provided by the tool to each judge at the beginning of every assessment session.

Inter-Judge Agreement Both in 2006 and 2007 all samples were assessed by domain experts, but to find out how important it is to involve domain experts in the assessment part of the work was repeated

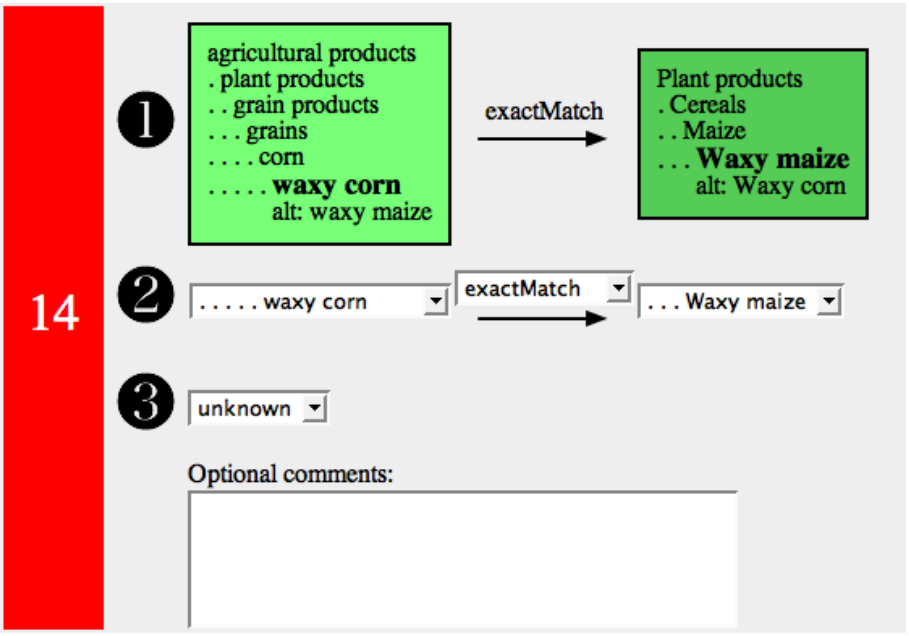

Fig. 6. Screengrab of the assessment tool used to evaluate Precision. Shown is the 14th mapping relation from a sample set of mappings, nalt:'waxy corn' skosmap:exactMatch agrovoc:'Waxy maize'.

2006		NAL & FAO (KR and food experts)		
		true	**false**	**unknown**
computer science	**true**	253	13	0
researchers at EKAW	**false**	6	52	0
(KR experts, agriculture laymen)	**unknown**	4	8	0

Table 3

Comparison between the assessments by judges from the NAL and FAO and computer scientist judges. Shown is a confusion matrix of the 336 alignments from the OAEI 2006 food task that were judged by both groups. Each alignment was judged once by someone from each group.

by laymen, computer scientists at the EKAW workshop (mainly knowledge representation experts). The agreement between the group of domain experts and the group of computer scientists was 72%. The computer scientists were less likely to judge a mapping to be correct than the domain experts. They judged 78% of the sample mappings to be "true", while the domain experts judged 85% to be "true". A more exact analysis is shown in table 3, which shows the judgements of the overlapping set of alignment relations. From this data we can compute Cohen's kappa to show how similar the judgements of the two parties are. We use Cohen's kappa as opposed to, for example, Fleiss' kappa, because we only have one judgement for each alignment per group and thus only two parties that can agree or disagree. Cohen's kappa is defined as follows:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the probability that agreement is due to chance. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$. A detailed description can be found in (Cohen, 1960). The κ among the two groups of judges was 0.73, which signifies a substantial agreement, which is higher than we expected. A κ of around 0.65 is not unusual between domain experts that are supposed to agree. Apparently, most alignments that are true are clearly true and slightly less, but still many of the false alignments are clearly false. We questioned some of the judges from both groups. The laymen tended to be more sceptical about the correctness of an alignment relation, because they felt it was

worse to make an inappropriate generalization than an inappropriate discrimination, whenever they were not familiar with the kind of generalizations that are common in agricultural library systems. If we would have used the assessments made by the laymen for this evaluation instead of those made by the domain experts the estimated Precision scores would have been slightly lower, but it is unlikely that the ranking of the participants would have changed.

Significance Testing As a significance test on Precision scores of the systems we used the Bernoulli distribution (van Hage et al., 2007). Precision of system A, P_A, can be considered to be significantly greater than Precision of system B, P_B, if their estimated values, $\hat{P}_A$ and $\hat{P}_B$ are far enough apart. In general, based the Bernoulli distribution, this is the case when the following equation holds:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\frac{\hat{P}_A(1-\hat{P}_A)}{n_A} + \frac{\hat{P}_B(1-\hat{P}_B)}{n_B}} \tag{3}$$

where n_A and n_B are the size of the set of assessed alignment relations that were returned by respectively system A or B. This number is always less or equal to the numbers in table 2, which shows the total number of assessed relations for all systems. The significance test in Equation 3 was used to determine which of the systems performs best on each of the three or four strata. The greatest error margin occurs when both systems have a Precision close to 0.5, when it is at most $\frac{1}{\sqrt{n}}$. When the results of the strata are combined, we are able to distinguish smaller differences in the results than for each of the strata alone. The upper bound of the error is equal to the error of simple random sampling (Cochran, 1977). The significance test we used for the combined result is as follows. We denote the estimated Precision of system A on stratum h as $\hat{P}_{A,h}$, the size of stratum h as N_h, and the size of the sample from stratum h as n_h (see table 2). We can conclude that system A performs significantly better than system B when the following equation holds:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\sum_{h=1}^{L} \frac{\hat{P}_{A,h}(1-\hat{P}_{A,h})}{N_A}\left(\frac{N_h}{n_h} - 1\right) + \sum_{h=1}^{L} \frac{\hat{P}_{B,h}(1-\hat{P}_{B,h})}{N_B}\left(\frac{N_h}{n_h} - 1\right)} \tag{4}$$

The greatest error margin still occurs when both systems have a Precision close to 0.5, but it is as most $\frac{2}{\sqrt{2n}}$. Again, the overall significantly best performance is indicated with a $\circ$ and if the best result was not significantly higher than the runner up this is indicated with a $\bullet$.

4.2. Recall

We estimate Recall using cluster sampling from the set of all *Correct* alignment relations. These are exactly all the relations that would be in an ideal finished alignment. This set is the same for all systems. It is the same for 2006 and 2007 with the exception of changed or added concepts. Therefore, we can use the same samples for the estimation of Recall for all systems. We also chose to reuse the samples we used in 2006 for 2007 with some updates. An advantage of this is that the results of 2006 and 2007 are easily comparable. A disadvantage is that there is the possibility that participants will train their systems on the samples and thus achieve better performance on the samples than on the rest of the alignment. This can cause a positive bias in the results. We were not concerned about this, because each participating system was only allowed one configuration for all the OAEI tasks. The food task is only one of all the tasks. The others focus on anatomy, directories, jobs, conferences, and dutch libraries. Each task has a different optimal setting for the ontology alignment systems. Therefore, specific optimization on the food Recall samples is unlikely, because it is probably counter productive for the participants.

Like the samples we used for Precision, the samples we used for the evaluation of Recall are on a certain topic. We chose several sub-hierarchies of the two thesauri and manually created the full alignment between the concepts in these selections. The topics we used in 2006 are: all *oak trees* (everything under the concept representing the Quercus genus), all *rodents* (everything under Rodentia), geographical concepts of *Europe* (countries), and everything under the NALT concept animal health and all AGROVOC concepts that have alignment relations to these concepts and their sub-concepts. The sizes of these samples

topic	size	% exactMatch	used in year **2006**	**2007**
animal health	34	57%	✓	✓
oak trees (taxonomical)	41	84%	✓	✓
rodents (vernacular)	42	32%	✓	✓
Europe (country level)	74	93%	✓	✓
topography (below country level)	164	35%		✓

Table 4
Sizes of the sets of manually created alignments used to evaluate Recall.

are shown in Table 4, along with the percentage of the alignment relations that was of type exactMatch, as opposed to broadMatch and narrowMatch. The average percentage of exactMatch in the 2006 sample was 70%. In 2007 we chose to add an additional geographical sample, *topography* below country level, because the 2006 geographical sample gave the impression that the percentage of exactMatch relations in the geographical domain is much higher than it really is. This is especially the case for concepts below country level, like provinces, which often do not have an exact counterpart in the other thesaurus and thus require a broadMatch or narrowMatch relation to be aligned.

Mapping Tool for Recall To create these samples we used the AIDA Thesaurus Browser, a SKOS browser that supports parallel browsing of two thesauri, concept search, mapping traversal, and the addition, change and removal of mappings of the SKOS Mapping Vocabulary. This tool was developed at TNO by Willem Robert van Hage in the context of the VL-e project.[14] It is an AJAX application that accesses a SOAP service wrapper of Sesame 1.2[15] through Java servlets. The service wrapper is part of the AIDA web service toolkit, which also includes wrappers for the Lucene search engine and several machine learning tools.[16] A screengrab of the tool is shown in figure 7. A preliminary version of the Recall samples was made at the Vrije Universiteit Amsterdam and was verified and extended by domain experts at the the FAO and USDA to produce the final Recall samples. The samples can be downloaded from `http://www.few.vu.nl/~wrvhage/oaei2006/gold_standard` and `http://www.few.vu.nl/~wrvhage/oaei2007/gold_standard`. The guidelines used to make the mapping were the following:

1. Starting from the side of AGROVOC, try to find a skosmap:exactMatch for every concept in the sample. If this is impossible, try to find a skosmap:narrowMatch or skosmap:broadMatch. Always choose the broader concept of these alignments as narrow as possible and the narrower concept as broad as possible.
2. Investigate the surrounding concepts of the target concept in NALT. If the surrounding concepts are still on the topic for the sample, try to map this concept back to AGROVOC using skosmap:exactMatch. If this is impossible, try to find a skosmap:narrowMatch or skosmap:broadMatch.

Significance Tests The sample selection procedure we chose, which is based on completely aligning subhierarchies where we expect many alignment relations, saved us a lot of time. This made it feasible to construct Recall samples. The downside of this is that the results are not fully generalizable to a greater population, because an assumption for generalization to the sample frame is random selection where each element gets an equal non-zero probability to be selected. Therefore, the application of significance measures would produce meaningless results.

Recall for all the systems is shown in table 6.

[14] `http://www.vl-e.nl`
[15] `http://openrdf.org`
[16] `http://www.adaptivedisclosure.org/aida-toolkit`

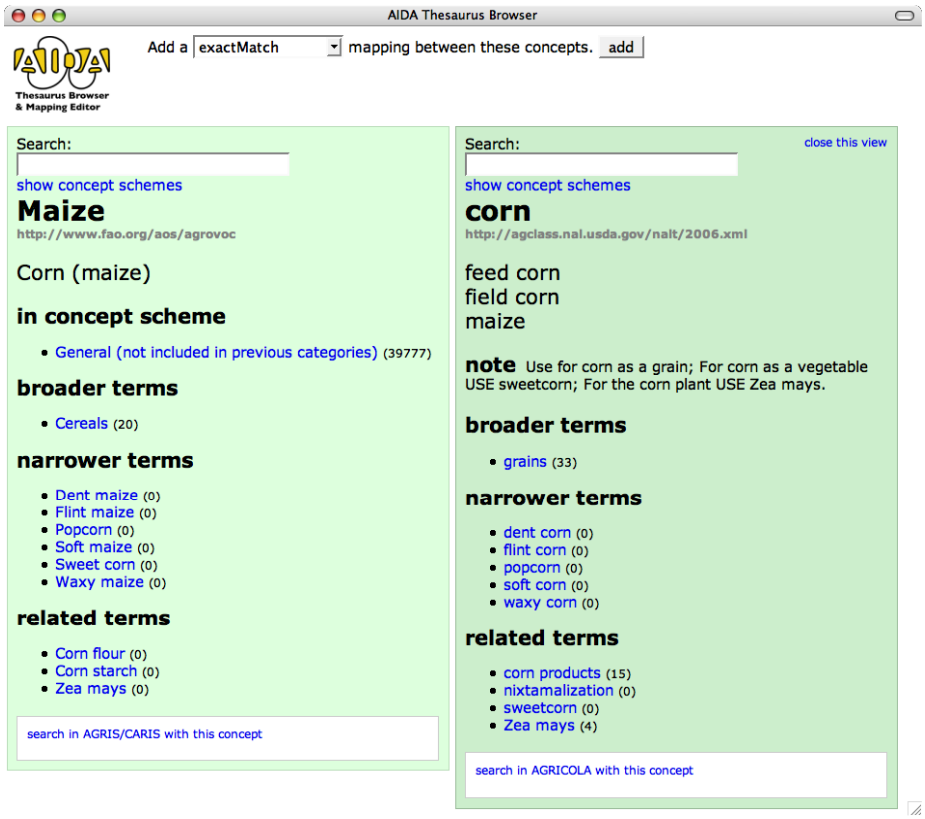

Fig. 7. Screengrab of the AIDA Thesaurus Browser, which was used to create alignment samples for the evaluation of Recall.

2006					
Precision for	RiMOM	Falcon-AO	Prior	COMA++	HMatch
taxonomical	0.82	0.83 ○	0.68	0.43	0.48
bio/chem	0.85 •	0.80	0.81	0.76	0.83
miscellaneous	0.78	0.83 ○	0.74	0.70	0.80
all topics	**0.81**	**0.83 •**	**0.71**	**0.54**	**0.61**

2007					SCARLET	
Precision for	RiMOM	Falcon-AO	X-SOM	DSSim	exact	broad & narrow
taxonomical	0.54	0.81 ○	0.26	0.37	0.60	0.13
bio/chem	0.84	0.91	0.92	0.86	1.00 •	0.17
geographical	0.97	0.95	1.00 •	0.94	0.00	1.00
miscellaneous	0.69	0.86 ○	0.62	0.57	0.75	0.44
all topics	**0.62**	**0.84 ○**	**0.45**	**0.49**	**0.66**	**0.25**

Table 5
Precision results based on sample evaluation.

5. Evaluation Outcome

In this section we will examine the quantitative evaluation results. We will first discuss the performance of the participating systems per year. Then we will look at the results of the systems that participated in both years (RiMOM and Falcon-AO) and investigate the difference. Finally, we will look into the performance of the systems' aggregated results.

2006

Recall for	RiMOM		Falcon-AO		Prior		COMA++		HMatch	
animal health	0.18	(0.55)	0.09	(0.27)	0.06	(0.18)	0.12	(0.36)	0.15	(0.45)
oak trees	0.85	(0.92)	0.83	(0.89)	0.85	(0.92)	0.54	(0.58)	0.80	(0.87)
rodents	0.07	(0.12)	0.00	(0.00)	0.05	(0.08)	0.02	(0.04)	0.00	(0.00)
Europe	0.70	(0.84)	0.69	(0.82)	0.65	(0.77)	0.24	(0.29)	0.68	(0.81)
all topics	**0.50**	**(0.71)**	**0.46**	**(0.65)**	**0.45**	**(0.64)**	**0.23**	**(0.33)**	**0.46**	**(0.65)**

2007

Recall for	RiMOM		Falcon-AO		X-SOM		DSSim		SCARLET all relation types	
animal health	0.21	(0.64)	0.21	(0.64)	0.00	(0.00)	0.06	(0.18)	0.00	(0.00)
oak trees	0.93	(1.00)	0.93	(1.00)	0.10	(0.12)	0.22	(0.24)	0.00	(0.00)
rodents	0.24	(0.42)	0.40	(0.71)	0.07	(0.10)	0.17	(0.29)	0.00	(0.00)
Europe	0.70	(0.84)	0.81	(0.97)	0.08	(0.10)	0.34	(0.40)	0.00	(0.00)
geography	0.26	(0.74)	0.32	(0.90)	0.05	(0.14)	0.18	(0.50)	0.01	(0.02)
all topics	**0.42**	**(0.78)**	**0.49**	**(0.90)**	**0.06**	**(0.11)**	**0.20**	**(0.37)**	**0.00**	**(0.00)**

Table 6

Tentative estimation of Recall based on sample evaluation. The numbers between parentheses show Recall when only the exactMatch alignments of the reference alignments are considered.

5.1. Results 2006

The Precision and Recall outcomes of the 2006 evaluation are shown at the top of Table 5 and Table 6. Overall, RiMOM and Falcon-AO were the best systems and COMA++ performed significantly worse than the other systems, mainly due to bad results for the taxonomical part of the task.

Precision The taxonomical parts of the thesauri accounted for by far the largest part of the mappings, 59% of all submitted mappings. The more difficult mappings that required lexical normalization, such as structure formulas, and relations that required background knowledge, such as many of the relations in the miscellaneous domain, accounted for a smaller part of the alignment. This caused systems that did well at the taxonomical mappings to have a great advantage over the other systems.

The RiMOM and Falcon-AO systems performed well at the largest two strata, taxonomical and miscellaneous, and thus achieved high Precision. What set them apart from the rest was mainly their strict acceptance criterion for alignments.

The COMA++ system lagged behind due to liberal lexical matching. Terms with a high edit distance were accepted as matches, for example, “Buttiauxella noackiae” and “Arca noae”, because both contain the substring “␣noa”. This was particularly harmful in the taxonomical stratum, because of three reasons. (1) Many latinized names have similar prefixes and suffixes and have a tendency to start with a ‘c’ or ‘p’. This decreases the edit distance amongst unrelated terms. (2) Different species from the same genus always share the same first name, for example “Camellia sinensis” and “Camellia japonica”. This greatly decreases the edit distance between different species. (3) It is not uncommon for species from completely different kingdoms, for example, plants and animals, to have the same specific name.[17] An example is “caerula”, which means blue and is contained in the scientific name of the blue tit (a bird), “Parus caeruleus”, and the blue passion flower (a flowering plant), “Passiflora caerulea”.

The HMatch system performed as well as the RiMOM and Falcon-AO systems, except in the taxonomical domain. This was due to the same reasons as those described previously for the COMA++ system, but on a smaller scale. Most of the mistakes for taxonomical alignment relations were due to point 2. Also, terms from completely different parts of the thesauri were matched when there was only lexical overlap. For example, “Jordan” (a river) and “Triglops jordani” (a fish).

[17] http://en.wikipedia.org/wiki/List_of_Latin_and_Greek_words_commonly_used_in_systematic_names

Recall All systems only returned skosmap:exactMatch mappings. This means Recall of all systems was limited to 71%. For example, RiMOM achieved 50% where it could achieve 71% and 71% where it could achieve 100% in table 6.

The RiMOM system managed to discover more good results than the Falcon-AO system on the four small sample Recall bases, at the cost of some Precision. These were mainly results where the preferred labels were different and had to be matched to an alternative label. For example, "Entomopathogenic fungi" and "Entomogenous fungi". RiMOM was less strict in these cases.

In general, performance on the rodents and animal health samples was bad. This was due to a large number of alignment relations in these sets that required some reasoning or background knowledge to find and a high number of broadMatch and narrowMatch relations. An example from the animal health set is the deduction that is required to conclude that "bee viruses" have a broadMatch "invertebrate viruses". A system will have to reason that bees are invertebrates. None of the systems was able to accomplish this. Many of the alignment relations from the rodents set required background knowledge, or reasoning over related term relations. In the NALT thesaurus the colloquial names of animals are linked to the scientific names with a related term relation. That means in order to match "Geomyidae" to "Gophers" it is necessary to recognize that this is a pattern in NALT.

The other sets, oak trees and Europe, were relatively easy for the systems. All systems except COMA++ were able to find around 70% to 80% of these alignment relations. There was no particular reason why the COMA++ system was unable to find a similar number of relations. The system simply returned only part of the results to boost Precision and selected the wrong part. For example, the match "Italy" and "Italy" was returned, but the match "Bulgaria" and "Bulgaria", which would have gotten at least the same confidence score, was not.

5.2. Results 2007

The Precision and Recall outcomes of the 2007 evaluation are shown at the bottom of Table 5 and Table 6. The RiMOM and Falcon-AO systems are still in the lead, but the RiMOM system showed a large drop in performance, while the Falcon-AO system seems to have made a small improvement over last year's results. The performance indications of SCARLET on the biological and chemical set looks higher than that of the other systems, but the total number of exactMatch relations SCARLET returned was only marginal. The number of relations returned in the biological and chemical set was only 2 and they were both correct. That means the best two systems on that set were X-SOM and Falcon-AO.

Precision The Falcon-AO system was clearly the best system in 2007. This was mainly due to its consistent behavior on the taxonomical set, but also the miscellaneous set. Other systems could match Falcon-AO on the biological and chemical, and geographical sets, but performed worse on the other two sets.

The X-SOM and DSSim systems show the largest difference in performance. The large majority of the taxonomical results can be attributed to extremely liberal use of edit distance matching without disambiguation using the structure of the thesauri. Many of these matches link concepts from completely unrelated parts of the thesauri. For example, "crushers" (equipment) has exactMatch "Mares" (animal). The only connection is that "crushers" has an alternative label "mashers", which also starts with "ma" and ends with an 's'. Another similar example is "housing" has exactMatch "Fomes" (a bracket fungus). The former concept has an alternative label "homes", which also ends in "omes". This phenomenon was the strongest in the taxonomical part, due to regularities in latin names described before.

Recall In 2007, the Falcon-AO system performed particularly well at the rodents set. There is an absolute difference of about 20% with the runner up, the RiMOM system. It is clear from the results that the context of the concepts, such as labels of related terms in NALT, are exploited whenever there is a lack of information. An example of a relation that was found is the "Geomyidae" to "Gophers" example, described before.

The X-SOM system had an unexpectedly low Recall on the Europe set. The X-SOM system has a string similarity module and the country names of the Europe set are lexically similar. However, it struggled with

the large size of the food thesauri. Therefore, we expect that the low Recall score is due to unfortunate partitioning of the thesauri, which precluded many matches from the result set.

The SCARLET system found almost none of the relations in the manually constructed alignments. Yet, a significant part of the relations that were returned were judged to be correct. The explanation for this paradoxical situation has to do with the evaluation method we used. The Recall samples consisted only of those relations that a human expert would create. These relations are all as strict as possible. Whenever an equivalent concept is available, an exactMatch relation is created. Only when no equivalent concept is available, a broadMatch or narrowMatch is created. These hierarchical relations are chosen as flat as possible, as explained in Section 4.2. All more diagonal relations can be inferred from these relations. For example, if there is no equivalent for the concept "car", it would be aligned to "motorized vehicle" and not to "vehicle". If "motorized vehicle" is a narrow term of "vehicle" then we can already deduce from that broader/narrower relation and the alignment relation that "car" also has a broadMatch "vehicle". Most of the relations that were found by the SCARLET system were very diagonal while a much flatter correct alignment relation was available. By our strict evaluation method, which measures how close the system's output is to human output and not how close their logical consequences are, nearly no correct relations were found. This is a pessimistic outcome. A more optimistic, and for some use cases perhaps a more realistic, outcome could have been calculated using the Semantic precision and Semantic recall metrics (Euzenat, 2007) instead of the Precision and Recall metrics we used.

5.3. Comparison 2006–2007

There were two systems that participated in 2006 and 2007, the RiMOM and Falcon-AO systems. The RiMOM system was changed considerably in the meantime, while the 2007 Falcon-AO system was simply an improved version of the 2006 Falcon-AO system.

Precision The RiMOM system had an unexpected fall in Precision from 2006 to 2007. This was due to bad performance in the taxonomical and miscellaneous sets. The main reason is that in 2007 the RiMOM system also returned many alignment relations that are based on partial lexical matching. Even though many of these partial matches are long or even complete words, for example, "fat substitutes" has exactMatch "Caviar substitutes", they are still more often incorrect than correct.

The Falcon-AO system showed a small drop in performance on the taxonomical test set, but a big improvement on the biological and chemical set. This was due to the decision to use edit distance instead of I-Sub for lexical similarity on the food task. I-Sub works better for short terms, while edit distance works better for long terms. Most terms in AGROVOC and NALT are quite long. Edit distance is more tolerant to small differences between terms than I-Sub. This allowed matches between chemical terms that only differed by a hyphen or a set of parentheses, like "parathion-methyl" to "Parathion methyl", which are common in chemical names. It also allowed matches between species names that are only subtly different, yet refer to completely different species, like "Helostomatidae" (a fish) to "Belostomatidae" (a beetle). In general, the Falcon-AO system performed better in 2007 than in 2006 due to improvements in the matching strategy. Apart from bug fixes, a big difference is that the more correspondences based on lexical matches with a high confidence are found the less hard the system try to find additional matches using less dependable matchers, such as its context matcher. This precluded many bad matches to be included in the result set when better lexical matches were already included. The results of this strategy are very similar to those of RiMOM's risk minimization strategy in 2006.

Recall There was a large Recall improvement by both RiMOM and Falcon-AO. Especially in the animal health and rodents sets. These were the harder sets to produce. Both systems employed a more tolerant lexical matching technique, which led to more matches. The Falcon-AO system was better capable of making the final decision which alignment relations to include in the result set than RiMOM. For example, the simple mapping of "Rats" with alternative label "Rattus" to "Rats", fell outside the final selection of results by RiMOM, but was returned by Falcon-AO.

	RiMOM 2006	Falcon-AO 2006	RiMOM 2007	Falcon-AO 2007
RiMOM 2006	1	0.75	0.48	**0.91**
Falcon-AO 2006		1	0.46	0.74
RiMOM 2007			1	0.50
Falcon-AO 2007				1

Table 7

Jaccard similarity ($|A \cap B| / |A \cup B|$) between the sets of submitted alignment relations of RiMOM and Falcon-AO in 2006 and 2007. The results of RiMOM 2006 and Falcon-AO 2007 are remarkably similar.

2006					
mapping found by # systems	1	2	3	4	5
average Precision	0.06	0.35	0.67	0.86	0.99
# mappings	21,663	2,592	2,470	4,467	5,555
2007					
mapping found by # systems	1	2	3	4	5
average Precision	0.19	0.81	0.88	0.91	–
# mappings	29,419	7,142	3,944	1,462	0

Table 8

Consensus: average Precision of the mappings returned by a number of systems.

System Similarity The results of the Falcon-AO 2007 system are very similar to those of the RiMOM 2006 system. They are even more similar to the results of the RiMOM 2006 system than to the Falcon-AO 2006 results. Table 7 shows the similarity between the sets of RiMOM and Falcon-AO for the years 2006 and 2007. The reason for this similarity is an improvement in Falcon-AO's lexical matching algorithm, which makes it very similar to that used by the RiMOM 2006 system. Most of the matches are derived mainly from evidence provided by lexical clues. The other matching strategies, such as Falcon-AO's GMO (structural similarity) or RiMOM's path similarity strategy, are minor sources of evidence. The RiMOM 2007 system focussed on adding extra sources of evidence, which hurt their performance, while the Falcon-AO system learnt of RiMOM's 2006 results and simply fixed the bugs in their lexical matching algorithm.

We expect that the overlapping part of the results of the Falcon-AO 2007 and RiMOM 2006 systems is close to the part of the alignment that can be acquired by means of lexical matching techniques and that the rest of the alignment can not be found using lexical matching techniques. Therefore, without the application of completely different sources of evidence, such as background knowledge in the form of third party ontologies or text mining, the performance of the Falcon-AO 2007 system is representative of the maximum performance one can expect for ontology alignment systems on thesaurus alignment tasks such as the food task.

5.4. Consensus

It has to be noted that a potential user of ontology-alignment systems does not necessarily have to limit himself to only one alignment system. Simple ensemble methods such as majority voting can improve Precision. To give an impression of this we list the average Precision of the alignment relations submitted by n systems in table 8. For $n = 4$ and 5 (*i.e.* the mappings that were returned by 4 out of 5 systems or all of the systems) Precision is significantly higher than for the best system by itself, Falcon-AO in this case. In 2006, nearly all of the 5,555 mappings found by majority voting are correct. Obviously, these are the "easy" mappings. Whether they are useful or not useful depends on the application of the mappings–if high Precision is more important than high Recall–and remains a topic for future research. In 2007 the result sets of the systems varied much more and thus majority voting worked less well, but still the quality of the alignment relations returned by 3 or more systems is well over that of the best system.

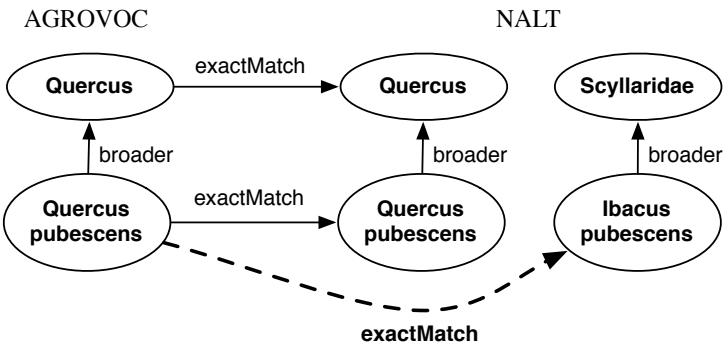

Fig. 8. Failing to recognize the naming scheme can lead to wrong mappings.

6. Analysis

In this section we will discuss a number of issues that limit the performance of alignment systems. Some of these issues are technical and are easy to solve. Others are more fundamental problems that cannot be solved soon if at all.

Inappropriate "spelling correction" Incorrect matches such as nalt:patients skosmap:exactMatch agrovoc:Patents and nalt:aesthetics skosmap:exactMatch agrovoc:anaesthetics are caused by inappropriate spelling correction code. In general, tolerating spelling differences in thesauri is not an effective technique, but if it is applied nonetheless it should only be applied when there is no exact literal match. For example, there is a concept representing "patients" in both thesauri. Recognizing this should trigger a alignment system to refrain from suggesting a mapping to "patents". The problem is greater for short terms than for long terms, because edit-distance based measures can be tuned better on long terms than on short terms, because the impact of changing a letter is greater in a short term than in a long term. Changing one letter in a short term changes its lexical shape more and is more likely to cause a difference in meaning than changing one letter in a long term.

Apart from incorrect partial phrase matches, like "disease reservoirs" to "water reservoirs", where a partial word overlap is assumed to indicate equivalence, inappropriate spelling correction is the most common source of mistakes. However, especially in the chemical domain, spelling correction also causes a great Recall gain.

Spelling correction should only be applied when the resulting term does not have a distinctly different meaning. A heuristic that could possibly be used to predict this is the comparison of word frequency distributions of the local textual context of the terms in some suitable large corpus of text. Currently, none of the ontology alignment systems implement this technique.

Labels following naming schemes Labels often follow naming schemes. Real-life ontologies often use more than one naming scheme. Both AGROVOC and NALT have a large section on biological taxonomy. The labels of these concepts follow the Linnaeic system of species names. Concepts in other sections of the thesauri (e.g. the sections on geography) follow different schemes. It is vital that lexical matchers recognize that different naming schemes require different matching rules. Perhaps the most common matching rule is postfix matching. This rule states that terms that end in the same word have similar meaning. For instance, "lime stone" and "sand stone" are similar. They are both kinds of "stone". Two terms from the Linnaeic system that end in the same word, such as "Quercus pubescens" (a tree) and "Ibacus pubescens" (a crustacean) are completely dissimilar. Failing to recognize that the Linnaeic system needs prefix matching and not postfix matching can lead to many wrong mappings. The bold arrow in figure 8 indicates this wrong mapping.

USE and USE FOR modeled with skos:altLabel When USE is modeled using skos:altLabel the difference between synonyms, obsolete terms, and acknowledgment of lack of detail disappears. In figure 9 AGROVOC does not include detailed descriptors for the concept nalt:Sigmodon. In fact, a few levels of

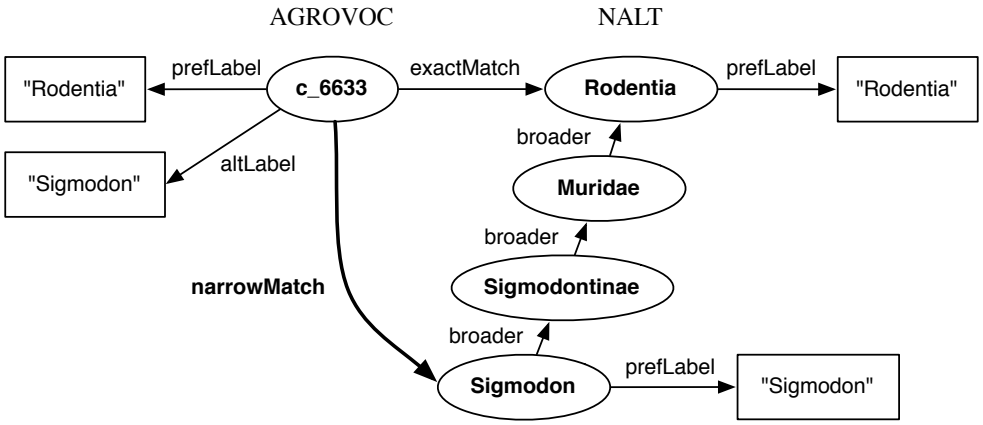

Fig. 9. USE modeled with skos:altLabel in AGROVOC.

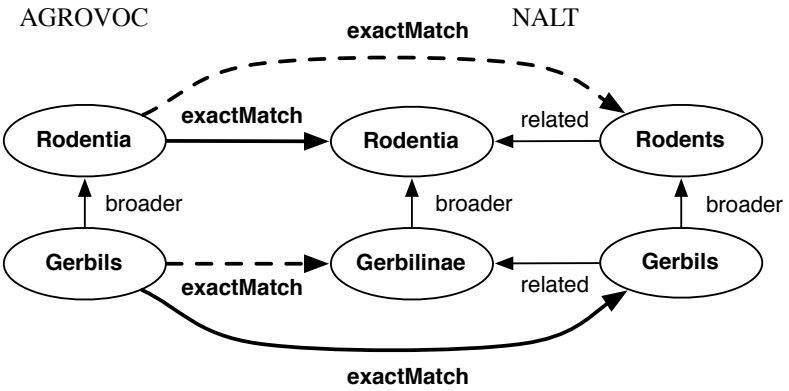

Fig. 10. Separate hierarchies for colloquial names and scientific names.

taxonomical distinctions are left out. The skos:altLabel "Sigmodon" is added to indicate this omission. It indicates that users that desire to refer to sigmodons should use the agrovoc:c_6633 concept, that symbolizes all rodents. A computer without prior knowledge about this modeling decision cannot distinguish this from synonymy represented with skos:altLabel. This will cause most systems to conclude there is a skosmap:exactMatch between agrovoc:c_6633 and nalt:Sigmodon, while the proper mapping between these concepts is a skosmap:narrowMatch.

Colloquial names and scientific names A delicate problem is that of colloquial versus scientific names for the same species. Take the example illustrated in figure 10 of gerbils with the scientific name "Gerbilinae". In NALT, the two types of names each have their own hierarchy, because colloquial names often do not exactly correspond to scientific names. There are Gerbilinae that are not Gerbils (*e.g.* sand rats and jirds), but there is no scientific name for Gerbils. It is also common to have scientific groups that have no colloquial name (*e.g.* nearly all taxonomical terms about bacteria). In AGROVOC the two are combined, because in the indexed documents they both refer to the same actual species. For example, "Roe deer" BT "Cervidae" BT "Ruminants". A complicating factor is indexing rules. In the AGRIS and AGRICOLA literature reference databases documents are indexed with scientific names whenever the animal or plant in the wild is meant, but the colloquial name is used when the domesticated animal or the product derived from the plant is. For example, "cows" are domesticated cows, while "Bos taurus" are wild cows, and "Zea mays" is the corn plant, while "maize" or "corn" is used for the product. The separation of colloquial and scientific names has the advantage that it enables more specific querying of the database, but that query expansion is necessary to find everything about cows or corn. Unification of colloquial and scientific names makes that easier, but makes finding things specifically about the product harder.

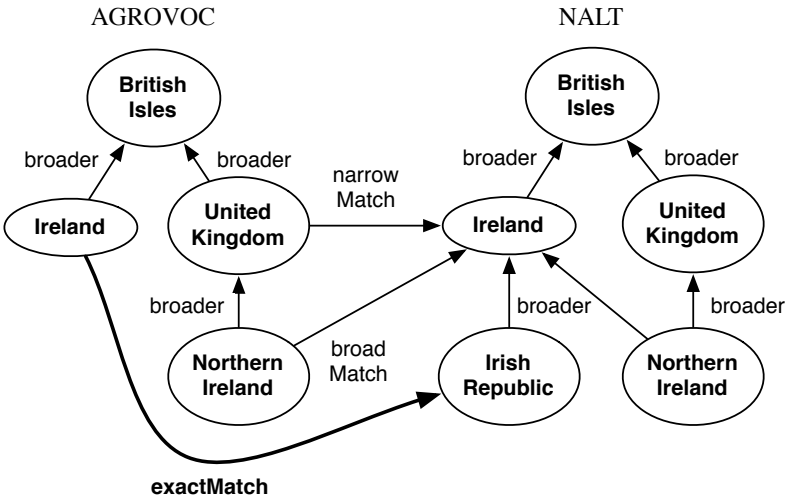

Fig. 11. Concepts representing different senses of a term.

Whenever alignment relations are traversed, it is clear that one enters another party's view of the world. The main reason to apply alignment relations is liberal query expansion. This considered, we are lead to believe that in the case of Gerbils there should be skosmap:exactMatch mappings to both hierarchies in NALT. We created the evaluation samples for Recall based on this assumption. Whether it is the proper treatment depends on the application of the mappings. Depending on the specific indexing rules of the collections, terms can symbolize different views of the concepts or refer to the same extension. This is not limited to species names, but also occurs with, for example, structural formulas of chemicals.

In AGROVOC and NALT this problem is extremely common, because the largest part of the thesauri deals with species names.

Clashing senses Sometimes all is not what it seems. Seemingly obvious mappings can be wrong. Consider "Ireland" and the "British Isles" in Figure 11. The British Isles can be partitioned in two ways, administrative and geographical. Respectively, the Irish Republic and the United Kingdom; or Ireland and the other islands of the British Isles, which all belong to the United Kingdom.

A natural intuition of people is the assumption that sibling concepts are disjoint. Therefore, if the distinction is made between Ireland and the United Kingdom, the most obvious interpretation is the administrative case. Even though in itself the word "Ireland" is more likely to refer to the island that to the nation, which is officially named the "Irish republic", people will immediately default to the nation. The lack of a broader relation between agrovoc:Northern Ireland and agrovoc:Ireland further supports their choice. Another natural intuition is that narrower concepts are strictly narrower than (*i.e.* not equivalent to) their parents. This means that the existence of the concept nalt:Irish Republic makes people assume that nalt:Ireland refers to the entire island. The narrower concept nalt:Northern Ireland confirms this. In the example this means that agrovoc:Ireland should be equivalent to nalt:Irish Republic.

In this case, a computer could solve this problem if a few OWL statements were added that proclaim siblings to be disjoint and broader concept to be not equivalent to narrower concepts. This kind of approach, however, is likely to cause more harm than good in the entire thesaurus. Thesaurus concepts are inherently vague and such a strict interpretation often causes unintentional inconsistencies. A technique that uses the added axioms as heuristics might be more suitable.

Obviously, the *Colloquial names and scientific names* issue, described previously, is also an example of clashing senses. Hence, this issue is a common phenomenon. There might not be as many geographical concepts as taxonomical concepts, but in applications geographical concepts are amongst the most commonly used concepts. Many fielded or facetted search clients support geographical selection of resources. Some data sets are better separated by nation (*e.g* legal documents), others are better served by a geographical separation (*e.g.* weather data).

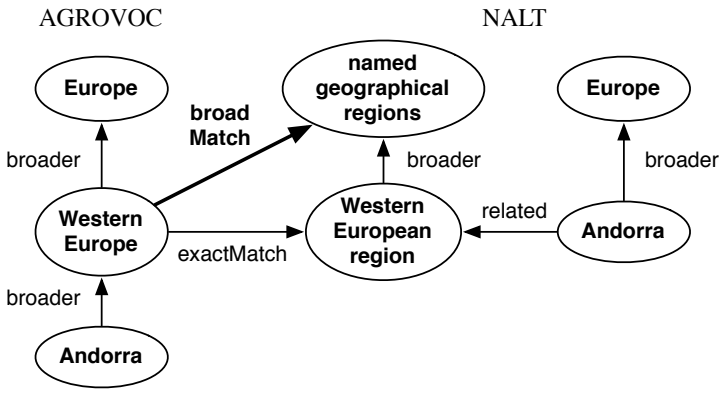

Fig. 12. The is no evidence in the thesauri for this skosmap:broadMatch.

No evidence in the thesauri for some correct mappings In many cases it is simply impossible to find certain mappings without resorting to external knowledge sources, such as a third ontology, concrete domain reasoning, text mining, or traditional knowledge acquisition. An example of a mapping that is impossible to find is shown in figure 12. Western Europe is clearly a named geographical region, but the skos:broader relation between nalt:Western European region and nalt:named geographical regions alone is not enough evidence to suggest this. AGROVOC contains no concepts that are lexically similar to the latter NALT concept.

An other example is: nalt:cytoplasmic polyhedrosis virus skosmap:broadMatch agrovoc:Reoviridae. None of the broader or narrower concepts have any lexical similarities, yet the mapping is sound. A search query on the MedLine collection with PubMed Central[18] reveals many articles that mention the relation. An excerpt from one of these articles that gives evidence for the mapping is: *"Cytoplasmic polyhedrosis viruses (CPVs) belong to the genus Cypovirus in the family Reoviridae (13, 36)."* (Ikeda et al., 2001)

This situation is common outside of areas where there is high consensus on the jargon (*e.g.* the taxonomical, geographical, or anotomical domains) and in the more general areas of the thesauri, *i.e.* near to the top concepts. In some areas (*cf.* the animal health Recall sample) alignments that require some degree of background knowledge are even the majority. The current ontology alignment systems, and even humans for that matter, have great difficulty to find these hard alignment relations. Therefore, the true magnitude of the problem is hard to quantify.

Useful broadMatch and narrowMatch are hard to find. The SCARLET system found thousands of hierarchical relations. A large part of these relations was correct, yet Recall scores on our samples are extremely low. This means that these relations are not the kind of relations domain experts would assert, even if many of them are not strictly false. Most of these relations are undercommitments. An example is the relation nalt:technology skosmap:narrowMatch agrovoc:Diesel engines. It appeared in the Precision sample for the SCARLET system and was judged to be true, but it would never appear in a manually constructed Recall sample. A thesaurus editor would always try to find the strictest relation that does not overcommit. AGROVOC has a concept agrovoc:Technology and NALT has a concept nalt:diesel engines. These two concepts would provide stricter matches and hence the alignment between nalt:technology and agrovoc:Diesel engines would never be asserted.

Whether undercommitments are a big issue depends on the application. If the only thing that matters for an application are the top concepts (*e.g.* for a rough separation of documents into topic categories) then undercommitments are no problem. If the alignment is used for delicate query expansion then undercommitments are nearly useless.

[18] http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=113995

7. Discussion

From this work we can draw conclusions on various levels: The specific challenges of thesaurus alignment in the agricultural domain, the importance of certain features for the quality of alignment systems in such tasks, the particularities of the evaluation of thesaurus alignment relations of various types, and in which cases thesaurus alignment can be automated with good results.

Specific challenges of the food task The main challenges for alignment systems in the OAEI 2006 and 2007 food task were the following:

- Compared to the data sets of the other OAEI tasks, AGROVOC and NALT are large. Only systems that could deal with the size of the AGROVOC and NAL thesauri (*e.g.* by correctly partitioning the data sets) could achieve satisfactory results. Some groups did not participate in the food task, because their systems were unable to load the thesauri. Most systems attempted some kind of partitioning. The quality of the partitioning turned out to be one of the decisive factors for overall system performance, for example the difference between Falcon-AO and X-SOM in Table 5 is partially caused by the different partitioning strategies of the systems.
- Only systems that were able to deal with the relatively weak semantic structure of thesauri could do well. Whereas most OWL ontologies have one label per class and a number of property restrictions, most SKOS thesauri have many labels, but lack property restrictions. This means systems could not rely on description-logic reasoning and were required to do term disambiguation. The systems that had the best lexical matching strategies (RiMOM 2006 and Falcon 2006 and 2007) performed significantly better than systems that focussed more on other facets.
- Both thesauri contain concepts from many different domains. Only systems that were able to do proper lexical analysis of labels that use various naming conventions could avoid large numbers of mistakes. Some systems did very well in some domains, but very poorly in other domains, for example, X-SOM did very well in the geographical and biochemical domain, but very poorly in the taxonomical domain. Systems that performed well overall were the clear winners.

Conclusions of the qualitative analysis The two most important conclusions of the qualitative analysis of the OAEI 2006 and 2007 food task results are:

- Within one thesaurus there can be many different kinds of labels (*e.g.* scientific names of species, structure formula's of molecules, named entities, medical terminology of various kinds, diverse types of jargon, etc.) Being able to deal with various naming schemes used in labels is, by far, the most important quality of a thesaurus alignment system.
- There is idiom in thesauri, "abuse" of the semantic features. For example, USE / skos:altLabel is sometimes used to indicate missing detail (see Section 6, page 18), that RT / skos:related usually also implies disjointness, and that BT / skos:broader should usually be considered as strictly broader. Alignment systems can gain by exploiting these rules.

Strict and relaxed evaluation methods For the evaluation of the alignments in the OAEI food task we chose to draw samples. Each sample alignment relation was either verified individually or constructed individually for the measurement of respectively Precision and Recall.

For the evaluation of broadMatch and narrowMatch relations there is a discrepancy between how we measured Precision and Recall. The correctness criterion for Precision could be summarized as: *"Is the relation valid?"*, while the criterion for Recall could be summarized as: *"Is the relation valid and as strict as possible?"*. The intuition behind the current Precision assessment criterion corresponds to that of Semantic precision, while the intuition behind the Recall criterion corresponds to strict Recall. We could have assessed Precision in the same strict way as we used for Recall or Recall in the same relaxed way as we used for Precision to get respectively a lower bound or an upper bound on the performance scores. This could have been accomplished by using either the current evaluation method for Precision and Semantic recall for Recall (relaxed), or a stricter criterion for Precision and the current evaluation method for Recall (strict).

The reason why we did not do this is a pragmatic one. We wanted to perform the exact same evaluation procedure for the food task in 2007 as we did in 2006. All of the systems in 2006 were only able to find exactMatch relations and for the evaluation of exactMatch relations there is no discrepancy between the current evaluation methods for Precision and Recall, because these relations are never an element of any other alignment relation's logical consequence. There are no equivalence relations in the logical consequence of an equivalence relation, only hierarchical relations.

A similar problem occurs in the evaluation of XML retrieval systems that are allowed to return nested parts of documents. These systems have to decide whether they should return the entire section, or only the most relevant paragraphs. A strict evaluation method states that only the most relevant elements (the smallest element containing all relevant information) should be returned. A relaxed evaluation method states that all enveloping elements or even contained elements can also be returned. The INEX evaluation (Kazai et al., 2004) initiative has experimented quite extensively with different combinations of strict and relaxed evaluation methods.

Application of thesaurus alignment Two important factors that determine how useful automatic ontology alignment can be in practice are the domains covered by the thesauri and the desired reliability of the results.

As we can see in Table 5 and 6, some domains are more easily aligned automatically than others. The geography domain, for instance, is an easy domain. The Falcon-AO 2007 system was able to find more than 90% of all exactMatch relations. Domains concerning roles of objects where there are different perspectives on the same objects are hard. An example is the category animal health (see Table 6) where you have mappings between, for instance, flukes as a species of worms and flukes as a kind of parasites. Or in the category rodents there are mappings between mice as a species and mice as a pest. The best systems were only able to find about 60% of the exactMatch relations and around 20% of all relations (see Table 6).

The fact that 90% of all equivalence relations between geographical terms can be found automatically by itself does not mean that it is always a wise decision to automate the alignment process for geographical terms. If you are dealing with an application where subtle differences are important, like the status of Northern Ireland or Montenegro, it is probably a better idea to construct the entire geographical alignment by hand. In many cases, this is feasible, considering the relatively small number of alignment relations in the geographical domain (as compared to, for example, the taxonomical domain). Judging by our experience with the OAEI 2006 and 2007 food task, we estimate that the verification of alignment relations consumes roughly 5 times less time than searching for the alignment relations by hand without suggested relations. So in some cases where Recall is high complete manual verification of an automatically-created alignment can potentially save time.

We presented a quantitative and qualitative evaluation of thesaurus-alignment techniques. Thesauri might be relatively weak semantic structures, yet they are widespread and used for a multitude of tasks in various contexts. This very versatility is what makes the evaluation of thesaurus alignment complicated. Ideally, every task gets its own evaluation method that takes into account its specific properties. For example, the evaluation of a classification task would use stricter measures than that of a browsing or recommendation task. As opposed to picking a number of different measures for different tasks we chose to pick a neutral evaluation measure. We complemented this quantitative evaluation with an in-depth qualitative analysis discussing the inherent strengths and weaknesses of the various alignment techniques employed by the systems.

Acknowledgements

We would like to thank the NAL and FAO for allowing us to use their thesauri and for the time and resources they committed to this work. Our special gratitude goes to everybody that helped with the assessment of the alignment samples: Gudrun Johannsen and Caterina Caracciolo at the FAO, Nicole

Koenderink, Hajo Rijgersberg, and Lars Hulzebos at the Wageningen Agricultural University, Fred van de Brug and Marco Bouman at TNO Quality of life, and Evangelos Alexopoulos at Unilever, and everybody at the EKAW 2006 workshop who took part in the inter-judge agreement experiment. Furthermore, we would like to thank the participants of the ECOTERM 2007 workshop for valuable discussions. This work was partly supported by the Dutch BSIK project Virtual Laboratory for e-science (`http://www.vl-e.nl`).

References

Castano, S., Ferrara, A., and Messa, G. (2006). Results of the hmatch ontology matchmaker in oaei 2006. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Clarke, S. G. D. (1996). Integrating thesauri in the agricultural sciences. In *Compatibility and Integration of Order Systems. Research Seminar Proceedings of the TIP/ISKO Meeting*.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons Ltd, 3 edition.

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

Curino, C. A., Orsi, G., and Tanca, L. (2007). X-som results for oaei 2007. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

Euzenat, J. (2004). An api for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference (ISWC 2004)*.

Euzenat, J. (2007). Semantic precision and recall for ontology alignment evaluation. In *Proc. of IJCAI 2007*, pages 348–353.

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Council of Library and Information Resources, report 91 edition. http://www.clir.org/pubs/abstract/pub91abst.html. ISBN 1-887334-76-9.

Hu, W., Cheng, G., Zheng, D., Zhong, X., and Qu, Y. (2006). The results of falcon-ao in the oaei 2006 campaign. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Hu, W., Zhao, Y., Li, D., Cheng, G., Wu, H., and Qu, Y. (2007). Falcon-ao: results for oaei 2007. In *roceedings of the International Workshop on Ontology Matching (OM-2007)*.

Huang, Z., van Harmelen, F., and ten Teije, A. (2005). Reasoning with inconsistent ontologies. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*.

Huang, Z., van Harmelen, F., and ten Teije, A. (2006). *Reasoning with Inconsistent Ontologies: Framework, Prototype and Experiment*, chapter 5. John Wiley & Sons Ltd.

Ikeda, K., Nagaoka, S., Winkler, S., Kotani, K., Yagi, H., Nakanishi, K., Miyajima, S., Kobayashi, J., and Mori, H. (2001). Molecular characterization of bombyx mori cytoplasmic polyhedrosis virus genome segment 4. *Journal of Virology*, 75:988–995.

Kazai, G., Lalmas, M., and de Vries, A. P. (2004). The overlap problem in content-oriented xml retrieval evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*.

Li, Y., Li, J., Zhang, D., and Tang, J. (2006). Result of ontology alignment with rimom at oaeiï£¡06. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Li, Y., Zhong, Q., Li, J., and Tang, J. (2007). Result of ontology alignment with rimom at oaei'07. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

Liang, A., Sini, M., Chang, C., Li, S., Lu, W., He, C., and Keizer, J. (2005). The mapping schema from chinese agricultural thesaurus to agrovoc. In *Proceedings of the fifth Conference of the European Federation for Information Technology in Agriculture, Food and Environment and the third World Congress on Computers in Agriculture and Natural Resources (EFITA/WCCA 2005)*.

Mao, M. and Peng, Y. (2006). Prior system: Results for oaei 2006. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Massmann, S., Engmann, D., and Rahm, E. (2006). Coma++: Results for the ontology alignment contest oaei 2006. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Miles, A. and Bechhofer, S. (2004). Skos simple knowledge organization system reference. http://www.w3.org/TR/skos-reference/.

Nagy, M., Vargas-Vera, M., and Motta, E. (2007). Dssim - managing uncertainty on the semantic web. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

Sabou, M., Garcia, J., Angeletou, S., d'Aquin, M., and Motta, E. (2007). Evaluating the semantic web: A task-based approach. In *Proceedings of the International Semantic Web Conference (ISWC 2007)*.

van Hage, W. R., Isaac, A., and Aleksovski, Z. (2007). Sample evaluation of ontology-matching systems. In *Proceedings of the 5th International Evaluation of Ontologies and Ontology-based Tools Workshop (EON 2007)*.