# Using quantitative aspects of alignment generation for argumentation on mappings

Antoine Isaac[1], Cássia Trojahn[2], Shenghui Wang[1], Paulo Quaresma[2]

[1] Vrije Universiteit, Department of Computer Science, Amsterdam, Netherlands
[2] University of Évora, Department of Informatics, Évora, Portugal

**Abstract.** State-of-the art mappers articulate several techniques using different sources of knowledge in an unified process. An important issue of ontology mapping is to find ways of choosing among many techniques and their variations, and then combining their results. For this, an innovative and promising option is to use frameworks dealing with arguments for or against correspondences. In this paper, we re-use an argumentation framework that considers the confidence levels of mapping arguments. We also propose new frameworks that use voting as a way to cope with various degrees of consensus among arguments. We compare these frameworks by evaluating their application to a range of individual mappers, in the context of a real-world library case.

## 1 Introduction

An important problem for ontology alignment is to find ways of choosing among the many tools and techniques available and their variations, and then combining their results. This is almost infeasible by purely manual efforts, and fixed heuristics for combining a pre-selected set of mappers will not fit a situation where more and more matching tools and options can be applied to an even greater variety of cases.

A first range of methods relies on (partial) evaluation of the results given by different techniques so as to *recommend* the best performing ones for the case at hand [1, 2]. Others anticipate such results by comparing the characteristics of the considered alignment case with "profiles" of matchers, as determined by previous evaluation [3]. However, these methods result in applying the same treatment to all the mappings obtained by a same method; they do not allow for considering each mapping. In the context of peer-to-peer systems, a more flexible approach has been proposed [4] that explores the way peers agree on a set of mappings, by evaluating the translations resulted from the application of each mapping when one peer queries for information provided by another.

A promising option is to use argumentation frameworks where arguments in favour or against mappings between concepts are declaratively represented and processed [5, 6]. Here, a set of mappers, representing different alignment approaches, generate a set of arguments that support the mappings. According to the definition of attacking relations, an argument for a mapping generated

by one mapper can be supported or attacked by other arguments from other mappers. Based on the framework instantiation (using specific attacking relation and preference order), it is possible to compute globally acceptable mappings.

These argumentation frameworks consider however the arguments based on their *intention* only. An argument against a concept mapping can successfully attack all the arguments in favour of it, even if there are dozens of these. In this paper, we investigate *quantitative* aspects of alignment generation among a set of arguing mappers. We focus especially on investigating and comparing the value, for the argumentation process, of alignment generation: (1) confidence level: can we use the confidence level of the mappings to solve argumentation conflicts? ; (2) consensus among mappers: can we use the agreement between mappers to measure the validity of the mappings in question?

In this paper, we re-use an argumentation framework that considers the confidence levels of mapping arguments [5]. We also propose new frameworks that use voting as a way to cope with various degrees of support for arguments. We compare these frameworks by evaluating their application to a range of state-of-the-art individual mappers, in the context of a real-world library case.

## 2 Argumentation Frameworks

The framework we have re-used and extended to deal with consensus, S-VAF, is based on Value-based Argumentation, itself based on Dung's classical system. In this section we present these three frameworks, as well as our new proposals.

### 2.1 Classical argumentation framework

Dung, observing that the core notion of argumentation lies in the opposition between arguments and counter-arguments, defines an argumentation framework (AF) as follows:

**Def.** [7] An Argumentation Framework is a pair $AF = (AR, attacks)$, $AR$ is a set of arguments and $attacks$ is a binary relation on $AR$.

$attacks(a, b)$ means that the argument $a$ attacks the argument $b$. A set of arguments $S$ attacks an argument $b$ if $b$ is attacked by an argument in $S$. The key question about the framework is whether a given argument $a \in AR$ should be accepted or not. Dung proposes that an argument should be accepted only if every attack on it is rebutted by an accepted argument. This notion then leads to the definition of acceptability (for an argument), admissibility (for a set of arguments) and preferred extension:

**Def.** [7] An argument $a \in AR$ is *acceptable* with respect to set arguments $S$, noted $acceptable(a, S)$, if $\forall x \in AR\,(attacks(x, a) \longrightarrow \exists y \in S, attacks(y, x))$

**Def.** [7] A set $S$ of arguments is *conflict-free* if $\neg\, \exists x, y \in S$, $attacks(x, y)$. A conflict-free set of arguments $S$ is *admissible* if $\forall x \in S$, $acceptable(x, S)$. A set of arguments $S$ is a *preferred extension* if it is a maximal (with respect to set inclusion) admissible set of $AR$.

A preferred extension represents a consistent position within $AF$, which defends itself against all attacks and cannot be extended without raising conflicts.

### 2.2 Value-based argumentation framework

In Dung's framework, all arguments have equal strength, and attacks always succeed, except if the attacking argument is otherwise defeated. However, as noted in [8], in many domains, including ontology alignment, arguments may provide reasons which may be more or less persuasive. Moreover, their persuasiveness may vary according to their audience. Bench-Capon has extended the notion of AF so as to associate arguments with the social values they advance:

**Def.** [9] A Value-based Argumentation Framework (VAF) is a 5-tuple $VAF = (AR, attacks, V, val, P)$ where $(AR, attacks)$ is an argumentation framework, $V$ is a nonempty set of values, $val$ is a function which maps elements of $AR$ to elements of $V$ and $P$ is a set of possible audiences.

Practically, in [6], the role of value is played by the types of ontology match that ground the arguments, covering general categories of matching approaches: semantic, structural, terminological and extensional. We argue further — and will use later — that any kind of matching ground identified during a mapping process or any specific matching tools may give rise to a value. The only limitations are (i) a value can be identified and shared by a source of mapping arguments and the audience considering this information (ii) audiences can give preferences to the values. An extension to this framework, required for deploying argumentation processes, indeed allows to represent how audiences with different interests can grant preferences to specific values:

**Def.** [9] An Audience-specific Value-based Argumentation Framework (AVAF) is a 5-tuple $VAF_p = (AR, attacks, V, val, valpref_{aud})$ where $AR$, $attacks$, $V$ and $val$ are as for a VAF, $aud$ is an audience and $valpref_{aud}$ is a preference relation (transitive, irreflexive and asymmetric), $valpref_{aud} \subseteq V \times V$.

$valpref_{aud}(v_1, v_2)$ means that audience $aud$ prefers $v_1$ over $v_2$. Attacks are then deemed successful based on the preference ordering on the arguments' values. This leads to re-defining the notions seen previously:

**Def.** [9] An argument $a \in AR$ defeats an argument $b \in AR$ for audience $aud$, noted $defeats_{aud}(a, b)$, if and only if both $attacks(a, b)$ and not $valpref_{aud}(val(b), val(a))$. An argument $a \in AR$ is *acceptable* to audience $aud$ with respect to a set of arguments $S$, noted $acceptable_{aud}(a, S)$, if $\forall x \in AR$, $defeats_{aud}(x, a) \longrightarrow \exists y \in S$, $defeats_{aud}(y, x)$.

**Def.** [9] A set $S$ of arguments is *conflict-free* for audience $aud$ if $\forall x, y \in S$, $\neg attacks(x, y) \vee valpref_{aud}(val(y), val(x))$. A *conflict-free* set of arguments $S$ for $aud$ is *admissible* for $aud$ if $\forall x \in S$, $acceptable_{aud}(x, S)$. A set of arguments $S$ in the VAF is a *preferred extension* for audience $aud$ if it is a maximal admissible set (with respect to set inclusion) for $aud$.

In order to determine preferred extensions with respect to a value ordering promoted by distinct audiences, *objective* and *subjective* acceptance are defined:

**Def.** [9, 6] An argument $a \in AR$ is *subjectively acceptable* if and only if $a$ appears in the preferred extension for some specific audiences. An argument $a \in AR$ is *objectively acceptable* if and only if $a$ appears in the preferred extension for every specific audience.

### 2.3 Strength-based Argumentation Framework

Value-based argumentation acknowledges the importance of preferences when considering arguments. However, in the specific context of ontology alignment, an objection can still be raised about the lack of complete mechanisms for handling persuasiveness. Indeed, off-the-shelf matching tools very often provide a mapping with a measure that reflects the strength of the similarity between the two entities, or a more general confidence they have in the mapping – almost always it is provided without any detail allowing to distinguish between the two. These measures – we will use *strength* in the following – are usually derived from similarity assessments made during the alignment process, *e.g.* from edit distance measure between labels, or overlap measure between instance sets, as in [10]. They are therefore often based on objective grounds.

However, there is no objective theory nor even informal guidelines for determining such strengths. Using them to compare results from different mappers is therefore questionable especially because of potential scale mismatches. For example, a same strength of 0.8 may not correspond to the same level of confidence for two different mapper.

It is one of our goals to investigate whether considering strengths gives better results or not.[3] To this end, we adapt a formulation introduced in [11, 5] to consider the strength granted to mappings for determining attacks' success:

**Def.** A Strength and value-based Argumentation Framework (S-VAF) is a 6-tuple $(AR, attacks, V, val, P, str)$ where $(AR, attacks, V, val, P)$ is a value-based argumentation framework, and $str$ is a function which maps elements of $AR$ to real values from the interval $[0, 1]$, representing the *strength* of the argument. An audience-specific S-VAF is an S-VAF where the generic set of audiences is replaced by the definition of a specific $valpref_{aud}$ preference relation over V.

**Def.** In an audience-specific S-VAF, an argument $a \in AR$ defeats an argument $b \in AR$ for audience $aud$ if and only if $attacks(a, b) \wedge ( str(a) > str(b) \vee (str(a) = str(b) \wedge valpref_{aud}(val(a), val(b))) )$

In other words, for a given audience, an attack succeeds if the strength of the attacking argument is greater than the strength of the attacked one; or, if both arguments have equal strength, the attacked argument is not preferred over the attacking argument by the concerned audience. Similarly to what is done for VAFs, an argument is acceptable for a given audience $w.r.t$ a set of arguments if every argument defeating it is defeated by other members of the set. A set of arguments is conflict-free if no two members can defeat each other. Such a set is admissible for an audience if all its members are acceptable for this audience $w.r.t$ itself. A set of arguments is a preferred extension for an audience if it is a maximal admissible set for this audience.

[3] Note that as opposed to what is done [11, 5] this paper aims at experimenting with mappers that were developed prior to the experiment, and hence more likely to present strength mismatches.

### 2.4 Argumentation Frameworks with voting

The previously described frameworks capture the possible conflicts between mappers, and find a way to solve them. However, they still fail at rendering the fact that sources of mappings often agree on their results, and that this agreement can be meaningful. Some large-scale experiments involving several alignment tools – as the OAEI 2006 Food track campaign [12] – have indeed shown that the more often a mapping is agreed on, the more chances for it to be valid.

In the following, we adapt the S-VAF presented above to consider the level of consensus between the sources of the mappings, by introducing *voting* into the definition of successful attacks. We first describe the notion of *support* which enables arguments to be counted as defenders or co-attackers during an attack:

**Def.** A *Support-aware Framework* (Sup-VAF) is a 7-tuple *(AR, attacks, supports, V,val,P,str)* where $(AR, attacks, V, val, P, str)$ is a S-VAF, and *supports* and *attacks* are disjoint (reflexive) binary relations over $AR$.

The voting is used to determine whether an attack is successful or not. Our first proposal opts for a simple voting scheme, where the number of supporters decides for success — as done in the *plurality voting system*.

**Def.** In a *Simple plurality voting Sup-VAF* an argument $a \in AR$ $defeats_{aud}$ an argument $b \in AR$ for audience *aud* if and only if
$attacks(a,b) \quad \wedge \quad ( \quad |\{x|supports(x,a)\}| > |\{y|supports(y,b)\}| \quad \vee$
$(|\{x|supports(x,a)\}| = |\{y|supports(y,b)\}| \wedge valpref_{aud}(val(a), val(b))) \quad ).$

This voting mechanism is based on simple counting. In fact, as we have seen previously, mappers sometimes return mappings together with a confidence value. There are voting mechanisms which address this confidence information. The first and most elementary one would be to sum up the strengths of supporting arguments. However, as for the S-VAF, this would rely on the assumption that the strengths assigned by different mappers are similarly scaled, which as we have seen is debatable in practice.

One possible option is to consider rankings derived from those confidence levels. First, we rank arguments on a value basis. For a given value $v \in V$, we define a function $rank_v : AR \longrightarrow \mathbb{N}$ that enables to order all the arguments according to their strength. Practically we choose to count, for each arguments, the ones that have a lower confidence level: $rank_v(a) = |\{x \in AR | val(x) = v \wedge str(x) < str(a)\}|$. Notice that this "ranking" reflects a partial order, as it allows for ties (for mappings with a same strength). It however avoids turning to random ordering decisions, and allows for seamless ranking of arguments derived from mappings that were not given any strength, by just considering that these arguments have an infinitely low strength. Based on this ranking, it is possible to define a voting process inspired by the Borda count method, which is one the reference methods for aggregating ranked choices – for each argument, we average the ranks given to it by the audiences which support it: [13]:

**Def.** In a *Borda count Sup-VAF* an argument $a \in AR$ $defeats_{aud}$ an argument $b \in AR$ for audience *aud* if and only if

$attacks(a,b) \wedge (\ bordaCount(a) > bordaCount(b) \vee$
$(\ bordaCount(a) = bordaCount(b)\ \wedge\ valpref_{aud}(val(a), val(b))\ )\ ),$
where $$bordaCount(arg) = \frac{\sum_{\{x|supports(x,arg)\}} rank_{val(x)}(x)}{|\{x|supports(x,arg)\}|}.$$

## 3 Experiments

### 3.1 Experiment case

Our testbed reproduces the Library Track of the 2007 OAEI campaign.[4] The National Library of the Netherlands maintains two book collections, each annotated with one thesaurus – *GTT* (35K concepts) and *Brinkman* (5K). These thesauri have to be aligned with links that correspond to classical thesaurus relations (*broadMatch*, *narrowMatch*, *relatedMatch*) or to semantic equivalence (*exactMatch*). It is important to mention that among the 2.4 Million of books in the two collections, 250K are actually dually annotated by both thesauri.

### 3.2 Mappers used

To carry out our experiments, we have selected the results of six mappers, which we believe to be a realistic sample of the available technology. The first three are state-of-the-art mappers developed by the community (OAEI participants), while the others result from our previous work. They exhibit a balance between generic methods – *e.g.*, string edit distance – and strategies that are arguably more appropriate to the case at hand – *e.g.*, using Dutch lexical knowledge.

*OAEI participants.* The first group of mappers we used are the participants of the OAEI Library Track: **Falcon** [14], **DSSim** [15] and **Silas** [16]. These tools are *hybrid*, as they use several alignment techniques in an integrated process. For instance, Falcon considers the similarity of both lexical and structural information of concepts, while Silas combines lexical techniques with applying instance-based similarity measures on books descriptions accessed from a library service. Note that, as generic matchers, they mainly return equivalence (*exactMatch*) mappings, except Silas, which provides a significant number of *related* matches.

*"Homegrown" mappers.* We also re-used mappers developed for previous experiments. First, an **edit-distance lexical mapper** applies string similiarity to (tokenized) labels, resulting in various exact equivalent, broader, narrower and related weighted matches. Second, a **Dutch SKOS lexical mapper** outputs weighted equivalent and broader mappings, based on Dutch morphological knowledge, exploiting the different type of labels of concepts as represented in SKOS. Third, an **extensional mapper** exploits the simple co-occurrence of concepts in KB book annotations [10] to produce weighted equivalence links. For more details, see `http://www.few.vu.nl/~aisaac/om2008/mappers-om08.pdf`.

[4] `http://oaei.ontologymatching.org/2007/library`

### 3.3 Evaluation measures

We set our evaluation in a scenario where mappings are used to translate book annotations from one thesaurus to the other [17]. One mapping – it is of course possible to restrict the mappings by selecting only one kind of relation, for instance *exactMatch* – is considered as a translation rule, which translates one GTT concept into its corresponding Brinkman concept. All mappings which involve the same GTT concept are aggregated into a single rule.

To carry out our evaluation, we use the 250K dually annotated books we have mentioned as a golden standard. For one such book, if one of its GTT annotation concept has a translation rule, we consider this book can be *fired*. Each of its GTT annotation concepts is then translated into its Brinkman correspondence(s). The original Brinkman annotation is taken as a gold standard, which is used to measure the quality of the generated mappings.

We measure how many translated concepts are correct (precision), how many real Brinkman annotation concepts are missed (recall), and a Jaccard overlap as combined measure of these two:

$$P_a = \frac{\sum \frac{\#correct}{|B_t|}}{\#books\_fired}, \quad R_a = \frac{\sum \frac{\#correct}{|B_o|}}{\#all\_books}, \quad J_a = \frac{\sum \frac{\#correct}{|B_o \cup B_t|}}{\#all\_books}$$

where $\#correct$ is the number of translated Brinkman concepts actually used, $B_o$ and $B_t$ are the original and translated Brinkman annotation, respectively.

### 3.4 Argumentation settings

*Characterisation of mapping arguments and attacking relation.* All the mappers we used return correspondences in the form of $m = (e_1, e_2, s, r)$, where $e_1$ and $e_2$ are entities from the two ontologies, $s$ a confidence level, and $r$ a mapping relation — *exactMatch*, *broadMatch*, *narrowMatch* or *relatedMatch*. Following [6, 5], arguments were created from these correspondences, as 6-tuples $arg = (e_1, e_2, s, r, v, h)$ where $v$ denotes a value or type of mapping argument (here, the tool which created the mapping) and $h$ a support token (+ or −, depending on whether the argument supports the correspondence or not). An *attack* relationship holds between two arguments if these involve the same pair of concepts but exhibit opposite support tokens.

*Generating negative arguments.* Our problem is to define the arguments which are *against* a given correspondence. The results of most of the state-of-the-art tools must be interpreted as supporting correspondences; except in some formal approaches, there is no "negative mapping". [6] solves this by examining the features of the concepts, such as their label or position in the ontologies' structural network, and use OWL semantics to find whether agents argue for or against a correspondence. In practice, this complex process amounts to re-define a mapping step, as the strategy and material used are very similar to the ones exploited by the individual mappers. Here, we propose to experiment with two simpler strategies which do not require to investigate the alignment space again.

*Negative arguments as failure* (NAF). This basic strategy relies on the assumption that mappers return *complete* results. For every possible pair of concepts and mapping relation, we check whether a mapper outputs it. If not, this correspondence is considered to be at risk, and a negative argument is generated, with an arbitrary strength of 1. This assumption, at first sight quite bold, is nevertheless supported by the observation that most mappers try to provide as many mappings as possible, the amount of (equivalent) mapping pairs being comparable to the size of the smallest ontology aligned.

*Negative arguments based on relation disjointness* (NARD). The second strategy assumes that two different thesaurus-inspired mapping relations (*broadMatch*, *narrowMatch* or *relatedMatch*) cannot hold between a same pair of concepts – a usual consistency check for thesauri – and that such a relation cannot hold between two equivalent concepts. An argument is thus considered to attack another if they link the same two concepts with different mapping relations.

*Frameworks tested.* For our evaluation, we experimented with the following selection of framework and attack strategy settings:

*Baseline.* This consists of a single aggregation – *union* – of mappers' results into a single set of mappings.

*F1* (Strength-based, attacks based on relation disjointness). This setting corresponds to the S-VAF described in Section 2.3 with the NARD attack strategy. Two versions are explored: ($F1_{cont}$) adopting the confidence values produced by the mapper as the strength of the generated arguments; ($F1_{disc}$) applying a threshold (0.5) on the original confidence values to produce arguments with a discrete strength — 0 if the confidence level is below 0.5, 1 otherwise.

*F2* (Strength-based, attacks based on absent correspondences). This setting corresponds to an S-VAF with the NAF attack strategy. The same two alternatives as for the previous framework are explored ($F2_{cont}$ and $F2_{disc}$).

*F3* (Plurality voting-based, attacks based on absent correspondences). This setting combines the Sup-VAF framework of Section 2.4 with the NAF strategy.

*F4* (Borda count-based, attacks based on absent correspondences). This is the Borda count Sup-VAF framework of Section 2.4, applying the NAF strategy.

*Mapper configuration.* For all settings, three groupings are considered: (1) the three *OAEI* participants; (2) our three *Homegrown* matchers; (3) *All* matchers.

*Preference ordering.* For all settings, we create an audience for each mapper involved. We define a complete preference order by defining a default order that is adapted, for each audience, by lifting itself to first position: for *OAEI*, the default order is Falcon>Silas>DSSim, but for the Silas audience the order defined is Silas>Falcon>DSSim. The default for *Homegrown* is Co-occurrence>SKOS lexical>Edit-distance. For *All*, it is Falcon>Co-occurrence>SKOS lexical>Edit-distance>Silas> DSSim. This order, even though inspired by observing respective mappers' general performances, remains rather arbitrary. Crucially, it is also fixed: we did not aim at analyzing the influence of this factor in our experiment.

### 3.5 Results and discussion

Tables 1 and 2 show the results we obtained – *w.r.t.* evaluation measures and amount of obtained annotation translation rules – both for individual matchers and their combinations. For brevity, we show the results of evaluation only when using *all* types of mappings in order to produce rules. We also performed evaluation using only the *exactMatch* ones, but that did not bring significant changes, both for absolute and relative performances of matchers and frameworks.

| Mapper | #Rules | P-a | R-a | J-a | | Mapper | #Rules | P-a | R-a | J-a |
|---|---|---|---|---|---|---|---|---|---|---|
| DSSim | 9467 | 13.3 | 09.4 | 07.5 | | SKOS | 13207 | 40.9 | 43.1 | 0.29.9 |
| Falcon | 3618 | 52.5 | 36.6 | 30.7 | | Co-occurrence | 15742 | 13.6 | 79.5 | 12.7 |
| Silas | 9358 | 45.5 | 42.6 | 31.4 | | Edit distance | 20065 | 31.6 | 43.5 | 24.4 |

**Table 1.** Individual mappers (P-a, R-a and J-a are expressed as percentages)

| | | OAEI | | | | | Homegrown | | | | | All | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Setting | | #R | P-a | R-a | J-a | | #R | P-a | R-a | J-a | | #R | P-a | R-a | J-a |
| Baseline | | 16990 | 32.6 | 46.8 | 26.0 | | 37421 | 13.0 | 79.8 | 12.3 | | 45052 | 12.0 | 80.0 | 11.4 |
| $F1_{cont}$ | | 16800 | 32.6 | 46.8 | 26.0 | | 36492 | 12.8 | 74.6 | 12.0 | | 43017 | 11.6 | 71.5 | 10.9 |
| $F1_{disc}$ | | 16799 | 32.6 | 46.8 | 26.0 | | 36332 | 12.1 | 70.3 | 11.3 | | 41222 | 10.8 | 66.7 | 10.2 |
| $F2_{cont}$ | | 829 | 52.6 | 07.5 | 07.2 | | 5021 | 52.8 | 37.0 | 31.3 | | 835 | 53.3 | 07.0 | 06.8 |
| $F2_{disc}$ | | 828 | 52.6 | 07.5 | 07.2 | | 7346 | 50.0 | 37.3 | 31.0 | | 833 | 53.2 | 07.0 | 06.8 |
| F3 | | 2816 | 53.6 | 31.5 | 27.4 | | 11912 | 41.9 | 45.3 | 29.2 | | 26721 | 07.6 | 78.8 | 07.3 |
| F4 | | 16970 | 32.5 | 46.6 | 25.9 | | 37383 | 13.0 | 79.6 | 12.2 | | 836 | 53.3 | 07.1 | 06.9 |

**Table 2.** Argumentation on combined mappers (P-a, R-a and J-a are expressed as percentages)

One can first observe the great difference between F1 and F2 – F1 filtering out only a few mappings compared to the baseline. The NARD strategy actually does not result in the generation of many counter-arguments, causing final results similar to those of the union of matchers. This is especially true for OAEI matchers, which output almost only exactMatch mappings – Silas outputs relatedMatch links, but these seem to relate concepts not involved in exactMatch links, even considering Falcon and DSSim. Results vary more for the Homegrown and All combinations, as these include many mappings with different relations, as well as with different strengths, implying more (successful) attacks. Making strengths discrete seems to have muscled up some counter-arguments, leading to slightly stricter (but less efficient!) selection.

F2 is much more selective. When a counter-argument with strength 1 is generated for one matcher, it is likely to defeat the positive arguments issued by matchers with lesser preference. For a given audience, a selective matcher causes the removal, from the subjectively acceptable mappings, of many results from all matchers below him. When each audience privileges the arguments produced by the matcher it represents, this amounts to filter out from the objectively acceptable mappings all those beyond the intersection of mappings with strength

1. This of course implies an expected great increase in precision and a decrease in recall, compared to the union of results. This also makes the practical interest of NAF with such a strength and preference configuration quite low. And it suggests further experiments, with different preference order patterns and default strengths for counter-arguments. For the OAEI combination (as well as for All, which includes it), the intersection is very small (caused by DSSim missing a lot of good mappings) which causes recall to be dramatically low. For the Homegrown configuration, which combines much less stringent mappers, the intersection is larger, explaining an evolution for precision and recall which is more beneficial. Note that there is almost no difference between the continuous and discrete settings for OAEI and All configurations. For these, the OAEI mappings almost entirely dictate the intersection, and most of them already have a strength of 1 – out of Falcon's 3,697 mappings, only 20 have a strength lower than 1. For the Homegrown configuration the effect is opposite to the one obtained for F1: a number of mappings are now "saved", as their strength being discretized up to the one of counter-arguments. However, even if saved mappings are numerous, their consequence on evaluation results is not striking, arguably because of their involving infrequent concepts in the collection. These observations lead to the conclusion that anticipating the effect of making strengths discrete is difficult, without more precise knowledge on the content of alignments.

For OAEI, the severe selection caused by NAF is partly compensated in F4 because of our ranking strategy. Falcon outputs a smaller number of precise results, all of them with a strength of 1. All the good mappings are therefore not attackable: if DSSim produces an attack on one Falcon correspondence, the rank of the attacker is very likely to be lower than the rank of the attacked.

The results for homegrown mappers hint at F3 being the only one able to compensate for attacks on correct correspondences, if enough mappers vote for them. This is certainly true for the OAEI combination, where framework 3 has produced the best precision. This is due the fact that using such framework, it is possible to retrieve significant part of the intersection sets of all mappings, considering the selection of the mappings based on supporters. For example, if both Falcon and DSSim have a positive argument in favour a mapping, independently of the strength of a possible negative argument against the mapping from Silas, the mapping is acceptable. But yet this is not always done at the cost of recall. Even if F3 had worse recall than Silas, it obtains more resulting mappings than F2 with the same continuous setting.[5]

The same applies for the "homegrown" combination. F3 has a slightly lower recall than F2 with continuous strengths, but, again, better precision and Jaccard average than the baseline results, and by an even greater margin. Even when individual mappers return large sets of overlapping mappings, argumentation with voting appears to be more promising than simple union. The results for the last All combination however hint that this positive effect may disappear

[5] Note that our evaluation strategy computes precision on the basis of books for which alignment allows to compute new annotations; it is therefore possible to have a greater set of mappings with a better general precision.

when the number of combined mappers gets bigger, and their precision lower. When too many lax mappers are involved, it is possible that wrong mappings find enough supporters to remain undefeated – the combined influence of DSSim and the un-filtered co-occurrence matcher may be instrumental here.

## 4 Related work and conclusion

Many methods, such as in [1–3], articulate mappings on a source basis: all mappings from a given source are selected (or weighted, in a weighted sum aggregation system) at once. This can be compared to the preference relation over mapping sources that we use. However, our framework is more precise, since it considers every mapping individually. In this respect, the alignment argumentation frameworks of [9, 5, 6, 8], which we re-use and extend, relate to the efforts focusing on the logical soundness of alignments. As an example, [18, 19] investigates how to detect individual mappings which cause inconsistencies, considering both aligned ontologies and proposed alignments. However, these approaches, similarly to the way argumentation is done in [6], require full-fledged formal ontologies, which will lack in many applications.

Instead, we have experimented with counter-argument generation techniques which can be applied to a wider range of cases. Our proposal to consider the strength of mapping arguments – and the consensus about them – assumes that quantitative aspects of alignment can help to compensate for the lack of formal knowledge, in contexts such as our library case.

However, our results are somehow inconclusive *wrt.* our initial research questions on the benefits of using strengths and consensus in argumentation. In some cases performances are comparable to those of best individual matchers. This is a significant outcome, when the best performing matcher is not known in advance. Still, no framework manages to outperform baseline merging for every configuration. Worse, results point at complex phenomena that may be inherent to combining alignments resulting from very different strategies – confidence assignments, filtering of results... Further investigation is therefore necessary.

First, we will complete our experiments by considering negative arguments based on relation disjointness for the frameworks 3 and 4 and comparing our results with using the basic VAF framework. Beyond, the problem of negative argument generation needs more attention. In our type of application scenarios, we cannot turn to formalized reasoning as done in [6]. It would be still interesting to investigate techniques that take into account more semantic constraints than done in our current strategies, using for instance detection of mapping cycles, or equivalence mappings that relates one concept to two distinct ones. We might benefit here from the constraints specified in the latest SKOS developments [20].

Relevance feedback, as used in [4, 1–3], is also absent in our argumentation system, in which only abstract arguments are considered. A possible option could be to combine both approaches, and raise counter-arguments based on the evaluation – either directly by assessing a correspondence, or in an end-to-end way by studying its effects on the application at hand.

**Acknowledgements** Authors are supported by the EU Programme Alban for High Level Scholarships for Latin America, the EU eContentPlus project TELplus and the Dutch NWO programme CATCH (STITCH project).

## References

1. Tan, H., Lambrix, P.: A method for recommending ontology alignment strategies. In: 6th Intl. Semantic Web Conference (ISWC 2007), Busan, Korea (2007)
2. Ehrig, M., Staab, S., Sure, Y.: Bootstrapping ontology alignment methods with apfel. In: 4th Intl. Semantic Web Conference (ISWC 2005), Galway, Ireland (2005)
3. Mochol, M., Jentzsch, A., Euzenat, J.: Applying an analytic method for matching approach selection. In: Ontology Matching Workshop, ISWC 2006. (2006)
4. Aberer, K., Cudré-Mauroux, P., Hauswirth, M.: Start making sense: The chatty web approach for global semantic agreements. J. Web Semantics **1**(1) (2003)
5. dos Santos, C.T., Moraes, M.C., Quaresma, P., Vieira, R.: A cooperative approach for composite ontology mapping. Journal of Data Semantics **10** (2008) 237–263
6. Laera, L., Blacoe, I., Tamma, V., Payne, T.R., Euzenat, J., Bench-Capon, T.: Argumentation over ontology correspondences in mas. In: 6th Intl. Conference on Autonomous Agents and Multi-Agent Systems. (2007)
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n–person games. AI **77** (1995)
8. Laera, L., Tamma, V., Payne, T.R., Euzenat, J., Bench-Capon, T.: Reaching agreement over ontology alignments. In: ISWC 2006. (2006)
9. Bench-Capon, T.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation **13** (2003)
10. Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: An empirical study of instance-based ontology matching. In: ISWC 2007, Busan, Korea (2007)
11. dos Santos, C.T., Quaresma, P., Vieira, R.: An extended value-based argumentation framework for ontology mapping with confidence degrees. In: Argumentation in Multi-Agent Systems, 4th Intl. Workshop, Honolulu, HI, USA (2007)
12. Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Svab, O., Svatek, V., van Hage, W.R., Yatskevich, M.: Results of the ontology alignment evaluation initiative 2006. In: Ontology Matching Workshop, ISWC 2006. (2006)
13. de Borda, J.C.: Mémoire sur les elections au scrutin. Histoire de l'Acadmie Royale des Sciences (1781)
14. Hu, W., Zhao, Y., Li, D., Cheng, G., Wu, H., Qu, Y.: Falcon-AO: results for oaei 2007. In: Ontology Matching Workshop, ISWC 2007. (2007)
15. Nagy, M., Vargas-Vera, M., Motta, E.: DSSim – managing uncertainty on the semantic web. In: Ontology Matching Workshop, ISWC 2007. (2007)
16. Ossewaarde, R.: Simple library thesaurus alignment with SILAS. In: Second Intl. Workshop on Ontology Matching, ISWC 2007. (2007)
17. Isaac, A., Matthezing, H., van der Meij, L., Schlobach, S., Wang, S., Zinn, C.: Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case. In: ESWC 2008, Tenerife, Spain (2008)
18. Stuckenschmidt, H., van Harmelen, F., Serafini, L., Bouquet, P., Giunchiglia, F.: Using c-owl for the alignment and merging of medical ontologies. In: Formal Biomedical Knowledge Representation Workshop, KR 2004, Whistler, Canada (2004)
19. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Applying an analytic method for matching approach selection. In: Ontology Matching Workshop. (2006)
20. Miles, A., Bechhofer, S.: Skos reference. Technical report, W3C (January 25 2008)