An evaluation of the performance of three semantic background knowledge sources in comparative anatomy

Ernest A.A. van Ophuizen, Jack A.M. Leunissen

Laboratory of Bioinformatics, Wageningen University, Droevendaalsesteeg 1
6708 PB Wageningen, The Netherlands

Abstract

In this paper we evaluate the performance and usefulness of three semantic background knowledge sources for predicting synonymous anatomical terms across species boundaries. The reference sources under evaluation are UMLS, FMA-OBO and WordNet, which are applied to the anatomical ontologies of mouse and zebrafish. Our results show that the use of specialized knowledge sources leads to highly accurate predictions, verified through complete manual curation, which can be further improved by combining multiple of said sources. We argue that these three references complement each other in terms of granularity and specificity. From our results we conclude that these references can be used to create reliable ontology mappings with minimal human supervision.

1 Introduction

Over the years, the amount of abstracts in Medline has been growing exponentially. It has become impossible to read everything published in one's field. The availability of electronic journals has made it easier to at least maintain a general overview of developments, but it is still very time-consuming and labour-intensive. Text mining provides a means to automate the search for relevant information, but despite advances in Natural Language Processing (NLP) there is still a lot of headway to be made. The ambiguity of terms used to describe one's findings make automatic data retrieval particularly challenging.

In order to facilitate automatic retrieval, controlled vocabularies have been constructed by consortiums of domain experts. (MGI [1], ZFIN [2], FlyBase [3]) These vocabularies then grew into ontologies, containing not only semantic information about individual terms, but also of the relationships between them. While this was an important step towards standardizing descriptions within a particular organism, there remain discrepancies between species; in part because of different choices by the respective consortiums and partly due to inherent differences between the organisms under comparison. Because of this, information about homologous structures is potentially lost in translation.

To navigate between species-specific ontologies, roadmaps are needed. These are constructed through ontology matching. In the past, efforts have been made to do this manually, such as SAEL [4]. While this may yield high-quality concept mappings, there are major drawbacks to a manual approach, which all stem from the fact that it is very labour-intensive and therefore expensive. Clearly, a largely automatic approach is the preferred option. Bastian *et al.* [5] have recently presented their method for matching based solely on information available in the ontologies, as have Ghazvinian *et al.* [6] These methods can be applied to any pair of closely related ontologies, not restricted to the field of biology.

Another approach is to make predictions based on external reference sources, as notably demonstrated by Aleksovski [7], Jean-Mary [8] and Marquet [9]. Based on the particular

reference used, these predictions can be made for a certain domain; when looking for similarities between a car and a motorcycle, one shouldn't use a medical reference.

Even automatic matching requires human supervision to a certain extent. The algorithm needs to be evaluated by manually checking the predictions it produces, before it can be applied confidently to a dataset. The produced prediction also requires some curation, but to a significantly lesser extent than the manual method. Building something automatically and manually removing the errors, assuming the algorithm is sufficiently discriminative, is much more efficient than building something by hand and having to contemplate every decision without the big picture present. A combination of the manual and automatic approach by Bodenreider [10] compared mouse and human with the Unified Medical Language System [11] (UMLS®) as a lexical reference.

In this paper we evaluate the suitability of three knowledge sources as references for the matching of anatomical ontologies. These knowledge sources are the UMLS Metathesaurus® and the Foundational Model of Anatomy [12] in obo-format (FMA-OBO) [13] - both specialized in anatomy - as well as WordNet [14], a non-specialized source. We use these references to match anatomical terms from mouse and zebrafish.

There are three important aspects to our approach. Firstly, the predictions have been manually curated in their entirety. Secondly, the sources are evaluated not only on their stand-alone performance but also on their ability to support the predictions made by the other sources. Finally, it does not only take synonyms into account, but also takes stock of parent-child relations. These can provide valuable circumstantial support in a subsequent matching effort in situations where synonyms could not be found.

2 Materials and Methods

All original code was written in Java, through the Eclipse IDE. Source material for anatomical terms was taken from the MGI Adult Mouse Anatomy ontology [15] (dated 09/06/2006) and the ZFIN Zebrafish Anatomy and Development ontology [16] (v. 1.6). We selected Mouse and Zebrafish as our first pair, because they are important model organisms for medical research. Additionally, they are sufficiently different from each other to present us with a challenging test case, yet not so dissimilar as to have hardly any overlap at all. The semantic comparison was made with Wordnet (v. 2.1), the UMLS Metathesaurus (v.2007AA release download) and FMA-OBO (v. 0.1). MySQL 5.0.18 was used for data storage.

WordNet was searched through the API. UMLS was queried with SQL queries against a local installation. FMA-OBO was first converted into a MySQL database and subsequently queried.

The obo-format files of the ontologies were parsed and each turned into a thesaurus. Every thesaurus entry consists of the original ontology ID, followed by the term name and any synonyms provided in the ontology. Abbreviations of less than three characters were deleted; they had been found to occasionally cause nonsensical mappings.

2.1 Syntactic comparison

For the syntactic comparison, the identifiers were removed from the thesauri. Every term or synonym was put on a separate line in ascending alphabetical order. The files were converted to all-lower case. The two lists were then compared to each other using the UNIX `diff` command. A matching line is considered a 'hit'.

2.2 Semantic comparison

The thesauri were screened with WordNet, UMLS and FMA-OBO, hereafter referred to as 'references'. The results would be stored in a file that will be referred to as an association list. Per identifier, each term or synonym was submitted to the reference of choice, in the manner described above. In case it was not recognized, the search term would be included in the association list along with a *synonym* code marker. If the term was recognized by the reference, both the search term and the reference's internal IDs corresponding to any results would be stored, each with a *synonym* marker. These IDs would then be submitted back to the reference, now querying it for first-order (i.e. direct) parent and child terms. This would generate a related set of results, whose reference IDs were also stored, along with a code to signify what relationship exists to the original term. This relation is a combination of (parent|child) and (is-a|part-of|unspecified), as indicated by the reference. One entry would thus consist of an ontology ID, a value (either a reference ID or the original search term, as it occurred in the ontology) and a relationship code. Example: <MA:0000003, C0460002, 1>, where MA:0000003 is the ontology ID, C0460002 is a reference ID (from UMLS) and 1 is our own relationship code for synonym. These entries were all stored in a MySQL database, one table for every species.

Cross-species synonyms were recovered by querying this database; if a value (i.e. reference ID or search term) in one table (i.e. species) was identical to the value in another table and both their relationship codes were *synonym* they would be considered predicted synonyms.

Cross-species parent-child pairs were recovered with a different query: if a value in one table was identical to the value in another table and the relationship code of one was *synonym* and the relationship code of the other was not *synonym* (hence by necessity a parent or child) they would be considered predicted related terms.

In this manner - for each of the three references individually - a list of cross-species synonyms and a list of cross-species parent-child relations was created, always for the combination mouse-zebrafish. The predictions were then manually curated using mainly Wikipedia [17] articles on anatomical structure. Wikipedia has been used as a tool for manual validation before in similar work [7] and a famous study [18] has found Wikipedia to be comparably reliable for scientific information to the Encyclopaedia Britannica. If the information in Wikipedia was insufficient to make a decision, we used Google to find additional sources of information. If a prediction was considered correct, it was given a score of 1; if it was considered incorrect, it was given a score of 0. If no clear classification could be established, it was given a score of 0.5. The results of this evaluation are listed in the results section of this paper. We listed reliability scores for every category, although some contain very few hits. Any score based on less than 30 hits is marked with an asterisk.

In first-order relations, a prediction was deemed correct if there was a linear relation between the two terms and a separation of at least one generation. In some cases, the relation was technically correct, but at such a distance to be non-informative. Such predictions were labeled as approximate instead of correct. The distance threshold was determined by retrieving the relationship from UMLS, if possible. Two ancestrally related terms spanning more than five generations in UMLS would be considered non-informative. This is elaborated upon in the discussion section, sub-header *Granularity* (section 4.3).

Occasionally two terms are predicted to be related as both synonyms and a parent-child pair. An example of this is the link between MA:0000060 (*blood vessel*; syn: vasculature) and ZFA:0001079 (*vasculature*; syn: blood vessels; syn: circulatory system), a result of poor synonym choices. Whenever this occurs, these terms are treated as synonyms. This policy keeps in line with our guiding principle of choosing the shortest route between two terms.

After analyzing the results, nine terms were found to be disruptive to the prediction. This could be because terms were too general to be useful ('body parts'), had disruptive organism-specific synonyms ('inner ear', syn. 'ear'; fish have no external ears) or inaccurately chosen synonyms ('trunk', syn: 'body'). All associated results - including correct predictions - were removed, to accurately analyze the method itself.

A schematic overview of the semantic method is provided in figures 1a and 1b.

3 Results

3.1 Syntactic comparison

The mouse ontology consisted of 2703 terms. By adding all listed synonyms, this was expanded to 3040. The zebrafish ontology consisted of 1558 terms, expanded to 2090. These expanded files were compared, resulting in 154 matches. These were all awarded a 100% reliability score without analysis and constitute a best-case scenario for string matching. The actual number of matched concepts is likely lower, as synonyms will cause multiple hits between two concepts.

3.2 Semantic comparison

For every reference thesaurus, two predictions were made: synonyms and parent-child pairs, as described earlier. The results of these are not merged, because they are more informative separately. They will therefore be presented as such.

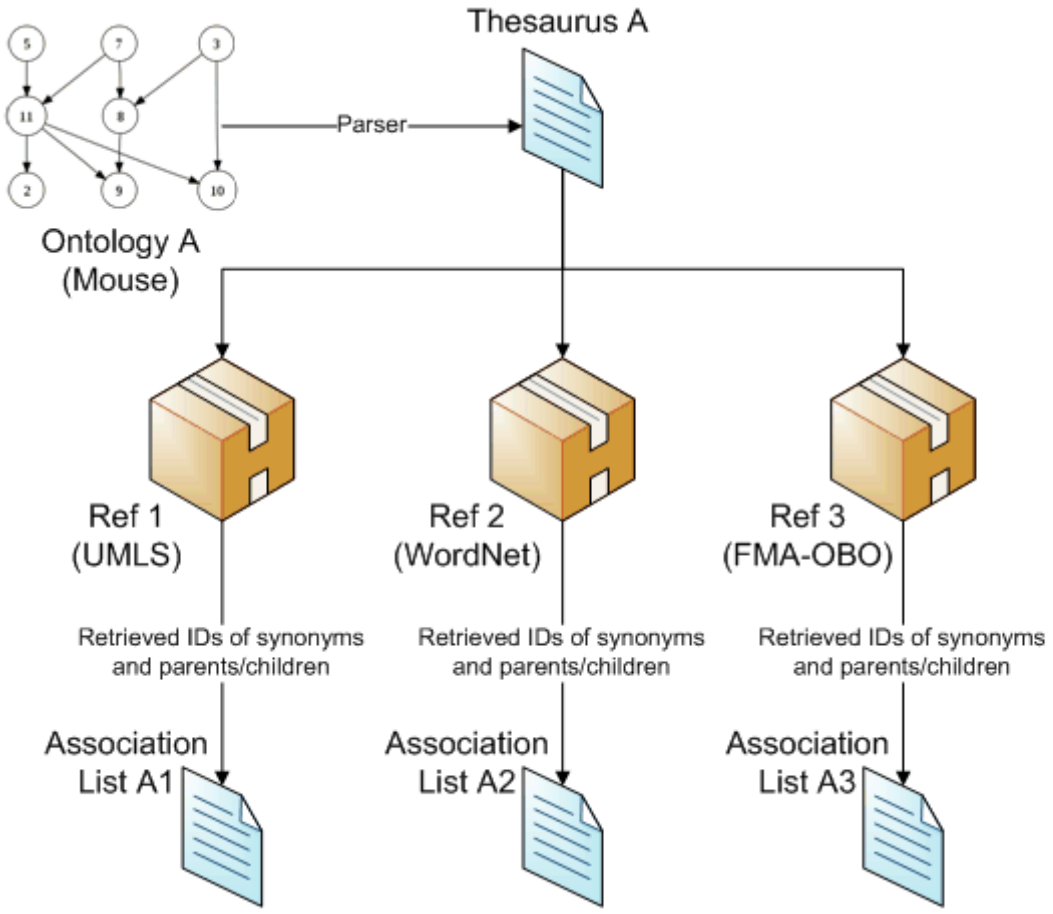

Figure 1a: An ontology is parsed into a thesaurus, which is then submitted to each of the references to produce reference-specific association lists

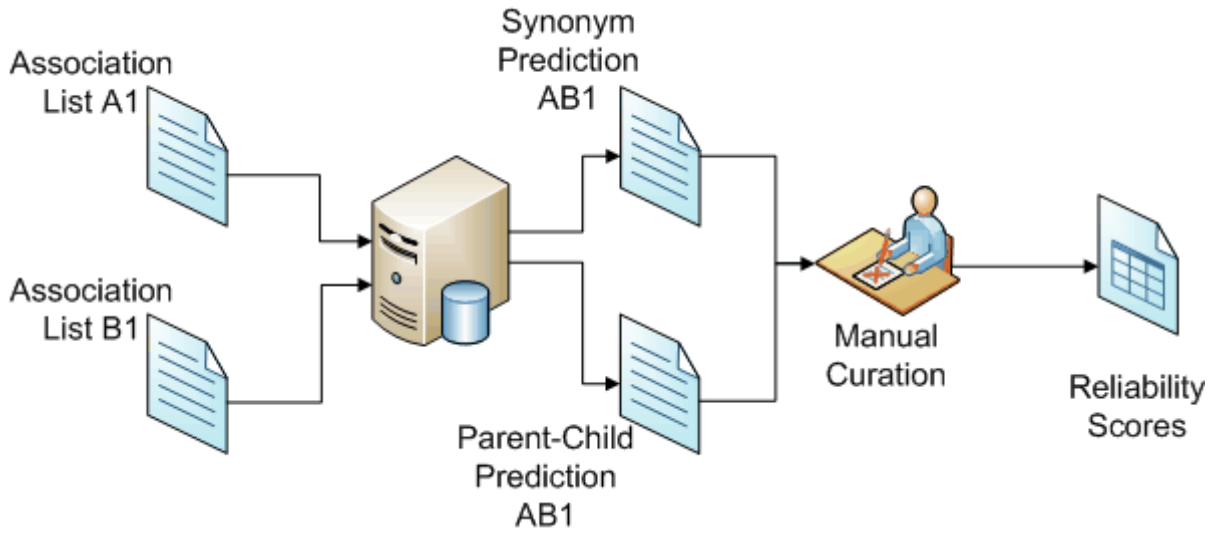

Figure 1b: Association lists from the same reference but different species are combined into synonym and parent-child predictions. After manual curation, reliability scores are calculated.

3.2.1 Synonym search

The automatic prediction using WordNet yielded 215 potential synonyms. Ten of these were discarded, being associated with one of the nine excluded terms. After manual curation of the remaining 205 hits, 161 of these were deemed correct and 28 false. The remaining 16 were not unequivocally correct or false. According to the scoring scheme, this gives WordNet a reliability of 82.4% for synonyms. A similar analysis was performed on the predictions from UMLS and FMA-OBO. (Table 1)

There is considerable overlap between the predictions from the three sources. Table 2 shows the amount of predictions supported by every combination of sources. The first three rows show that predictions stemming from a single source are not particularly reliable. The bottom row shows that those predictions supported by all three sources (62.8% of all predictions) are highly reliable. Table 2 is represented graphically as a Venn diagram in the supplementary materials.

When a term is mapped as a synonym for multiple terms in another ontology, ambiguity is created. Tables 3A and 3B show the amount of ambiguity arising from all three methods for both organisms.

3.2.2 Relatives search

The second step in the automatic prediction was that of so-called “first-order relatives”, i.e. parent-child pairs. WordNet analysis produced 426 unique predictions; 306 of them were correct, 51 were approximately correct and 69 were incorrect. UMLS produces many more predictions, which are on average more reliable. FMA-OBO also produces high-quality predictions, but in lesser number. A text comparison is unable to find semantic relations. (Table 4)

Table 5 is a breakdown of table 4’s predictions per source in the same way as table 2. It shows that about half of WordNet's predictions are not found with either of the other sources. Approximately half of those are correct and a third are completely wrong. The other two resources perform considerably better on their own, with at least 90% of their exclusive predictions correct. If a prediction is generated by multiple sources, it has a higher reliability score. Table 5 is represented graphically as a Venn diagram in the supplementary materials.

Table 1: Analysis of predicted synonyms

Method	predicted #hits	scratched	corrected #hits	correct	approx.	wrong	reliability
Text comparison	154	N/A	154	154	0	0	**100.0%**
WordNet	215	10	205	161	16	28	**82.4%**
UMLS	198	11	187	169	13	5	**93.9%**
FMA-OBO	165	9	156	151	4	1	**98.1%**

Table 2: Reliability by source. * denotes reliability scores in categories < 30 hits.

WordNet	UMLS	FMA-OBO	#hits	correct	approx	wrong	reliability
✓	-	-	44	7	9	28	**26.1%**
-	✓	-	20	9	6	5	**60.0%***
-	-	✓	3	2	0	1	**66.7%***
✓	✓	-	14	11	3	0	**89.3%***
✓	-	✓	0	0	0	0	**N/A**
-	✓	✓	6	6	0	0	**100.0%***
✓	✓	✓	147	143	4	0	**98.6%**

Table 3A: Ambiguity of predictions. Mouse terms mapped to 2 or more zebrafish terms.

Method	2	3	4	>4
WordNet	17	6	0	0
UMLS	9	2	0	0
FMA-OBO	8	0	0	0

Table 3B: Ambiguity of predictions. Zebrafish terms mapped to 2 or more mouse terms.

Method	**2**	**3**	**4**	**>4**
WordNet	20	5	3	1
UMLS	24	1	0	0
FMA-OBO	12	1	0	0

Table 4: Analysis of predicted first-order (i.e. parent-child) relations

Method	**predicted #hits**	**scratched**	**corrected #hits**	**correct**	**approx.**	**wrong**	**reliability**
Text comparison	N/A	N/A	N/A	N/A	N/A	N/A	N/A
WordNet	470	44	426	306	51	69	77.8%
UMLS	963	52	911	837	49	25	94.6%
FMA-OBO	234	17	217	201	13	3	95.6%

Table 5: Reliability by source. * denotes reliability scores in categories < 30 hits.

WordNet	**UMLS**	**FMA-OBO**	**#hits**	**correct**	**approx.**	**wrong**	**reliability**
✓	-	-	216	106	42	68	**58.8%**
-	✓	-	579	522	35	22	**93.2%**
-	-	✓	22	21	0	1	**95.5%***
✓	✓	-	137	133	3	1	**98.2%**
✓	-	✓	2	2	0	0	**100.0%***
-	✓	✓	122	113	7	2	**95.5%**
✓	✓	✓	71	67	4	0	**97.2%**

4 Discussion

4.1 Usability of the references

Each of the references used has its own strong points, which are suited more or less to the task at hand.

UMLS is a very fine-grained, exhaustive source of high-quality information. The fact that this information is tailored specifically to the field of medicine only increases its usefulness. It could be argued that the performance of UMLS solo in table 5 is somewhat padded. UMLS is the only source that contains the names of individual muscles. This accounts for 82 hits, where some muscle is correctly predicted as 'part of musculature system'. These are of course valid predictions, but if they are ignored, the score of UMLS solo drops to 92.1%. This is only a minor deterioration of 1.2% reliability.

WordNet is a more generalized resource, which can be observed clearly in the results. Ambiguity in natural language leads to strange predictions; ‘intestine’, ‘heart’ and ‘testicle’ obviously do not refer to the same anatomical concept. Their synonyms ‘guts’, ‘heart’ and ‘balls’ all refer to the concept of courage, though. Faulty synonyms propagate into the first-order relations, further worsening WordNet’s scores. Fortunately these problems are generally limited to widely-known structures; terms like ‘medulla oblongata’ are conspicuously absent from street slang. WordNet is still useful to our analysis; a prediction backed by both WordNet and UMLS is more reliable than a prediction based on UMLS alone. This is true for both synonym and first-order relation predictions. WordNet even spots a few synonyms that UMLS has missed, but this information is unfortunately lost in the noise when matching ontologies automatically.

FMA-OBO is fully incorporated in UMLS and hence shouldn’t add anything in the way of new predictions. This occasionally does happen, when FMA-OBO deems two concepts synonymous where UMLS ranks them as parent and child. Aside from this, FMA-OBO primarily adds a framework solidly rooted in formal ontological theory.

Based on the reliability values presented in tables 2 and 5, one can assess the likelihood of a certain prediction’s correctness. Predictions that are produced by multiple references are more reliable than those stemming from only a single reference. These highly reliable predictions can then be used as a strong framework to base the true ontology mapping on. This framework may then be used to provide proximity support for predictions of a lower reliability.

4.2 Ambiguity

The first thing one notices is that WordNet produces more and greater ambiguity than the other methods. This is a direct result from the ambiguity of natural language, as described in the previous paragraph. This kind of ambiguity is easily corrected, as UMLS and FMA-OBO are much less prone to mistakes of this kind. We are planning to construct a reliability model which also takes the amount of matching in the descendants into account. It is expected that such a model will reduce the detrimental effects of this type of ambiguity, as figurative speech usually doesn't propagate into sub-concepts.

A second major factor in ambiguity is caused by the anatomical ontologies themselves. If one ontology uses a term (e.g. *otic capsule*) as a synonym for two distinct concepts (*otic vesicle*, *otic capsule*), it is likely that a corresponding term in the other ontology will be matched to both concepts. This is unavoidable, but it is reasonable to assume that in these cases those two

concepts will be very close together in the ontology. In such cases, human intervention is necessary.

A third way for ambiguity to arise is a difference in opinion between the ontology and the reference. UMLS considers *carotid artery* and *common carotid artery* to be synonymous, whereas the mouse ontology lists them as separate terms. As a result, UMLS links both mouse terms to the zebrafish term *carotid artery*.

4.3 Granularity

The step size between parent and child is the granularity of a source. The most detailed descriptions indicate the depth. The granularity and depth of these resources differ greatly. One extreme is WordNet, which is both very coarse-grained and rather shallow. This can be expected, as WordNet is designed to be a semantic dictionary of the entire English language. Thus, the information density for any given field will be lower than in a specialistic source. The other extreme is UMLS, which is extremely fine-grained and very deep. This makes querying UMLS a rather expensive process in terms of processor time. When comparing a mouse to a fish, terms like *nail of left index finger* (C0926376) are a burden rather than a blessing. FMA-OBO holds a middle ground between these two.

This difference in granularity accounts for the fact that different sources find different first-order predictions. When one source considers something to be a single step, another may have several additional layers in between. Depending on how large that single step is, it may still be valuable information. To get an impression of the magnitude of this effect, we ran all the WordNet-exlusive parent-child predictions through UMLS and registered whether all predictions could be retrieved and if so, in how many steps. (Figure 2)

Regarding granularity, this graph shows that the WordNet-unique hits that are indeed retrievable can be found at a distance of 2 or 3 generations. Furthermore, two additional conclusions can be drawn. The first is that UMLS is unable to find certain terms which are found with WordNet (category NF) and that do lead to valid predictions. The second is that if UMLS returns no linear relation between the two terms (category NR), the prediction is probably not correct. We will use this information in our future automated mapping to further decrease the amount of human curation necessary.

4.4 Types of errors

An incorrect prediction can have a number of causes. One of these has been discussed earlier: ambiguity of natural language. When a word has figurative meanings, this may be the cause of incorrect predictions.

A second cause of errors is faulty synonym annotation. Here the fault actually lies with the ontology. Take the term *inner ear* from zebrafish (ZFA:0000217). Among its synonyms is ear. It stands to reason that if there's such a thing as an inner ear, there is bound to be an outer ear as well. And it is reasonable to assume that ear consists at least of said inner ear and outer ear. Most of the time this assumption holds, but in fish it obviously does not. Looking from a fish perspective, the terms ear and inner ear are indeed synonymous! The creators of the zebrafish ontology cannot really be blamed for these inconsistencies, but they do frustrate mapping efforts like ours. A very workable solution would be to create an ontology term ear in ZFA which has only one child: inner ear.

A third source of errors is analogy. Teeth are an example of analogy in the comparison between zebrafish and mouse. Mice have oral teeth, zebrafish have pharyngeal teeth. They are similar in appearance and function, but where oral teeth develop from the ectoderm, pharyngeal teeth develop from interactions between endoderm and mesenchyme. In a

generalized ontology, one could have a term *tooth* with two children: *oral tooth* and *pharyngeal tooth*, each with their own (though perhaps partly overlapping) progeny. However, within their respective species it is perfectly natural to simply talk about *tooth*.

Situations such as this can only be resolved through direct intervention of an expert. This is an expensive solution, yet unavoidable in cases such as this. The only alternative is a structural change in the source ontologies.

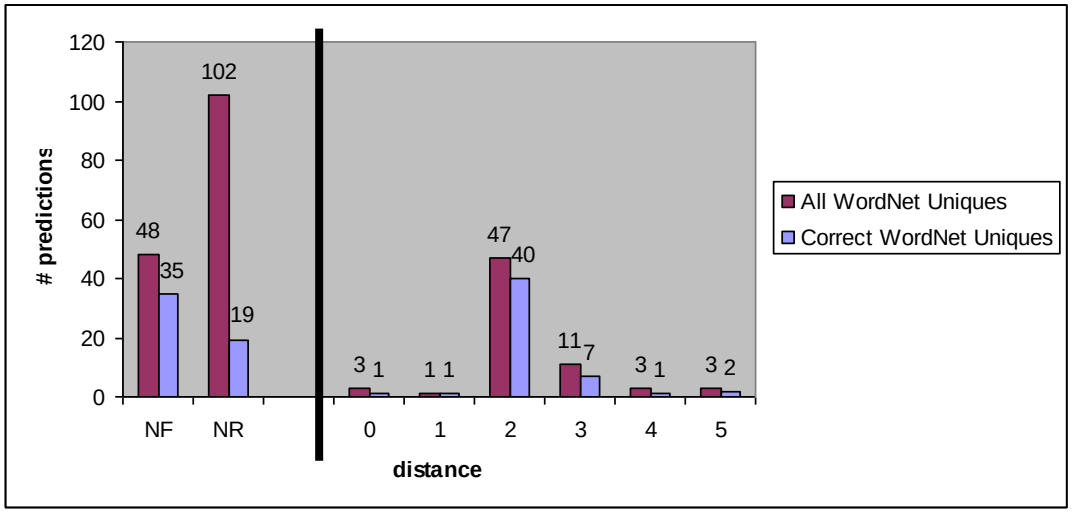

Figure 2: Distance at which WordNet-unique parent-child predictions were retrieved in UMLS. NF = Not Found, NR = Not Related

4.5 Curation with Wikipedia

As mentioned in section 2.2, manual curation of all predictions was performed using Wikipedia. To test the validity of this approach, a sample of these curations were verified with an anatomy textbook. [19] This sample consisted of 30 randomly selected predictions from the largest set (UMLS parent-child) which had previously been classified as 'correct'. Out of those 30, 28 were supported by the textbook. Of the remaining two, one could not be found, as *tegmentum* did not appear in the index. The only unsupported prediction was <reticular formation part_of hindbrain>. According to the textbook, it is part of the midbrain. The Wikipedia article states that it runs "through the mid-brain, pons and medulla", citing a different textbook. [20] The latter two are parts of the hindbrain. This cross-validation shows that Wikipedia is a reliable source for the purpose of curating predictions.

4.6 Adding more species

Now that we have evaluated these three references, it is time to add more species to the comparison. The parser accepts any ontology presented in obo-format; all it takes is to change the source and destination file and possibly the strings that the ontology uses for relations. Apart from Mouse and Zebrafish, there are anatomical ontologies for numerous species, ranging from nematodes and fruit flies to frogs and humans.

If an ontology of interest is unavailable in obo-format, it requires a different parser. That parser's output format should be the same as our original. The matching algorithm will then have no technical problems with generating predictions, as the essential components (ID and name) are present in every ontology.

5 Conclusion

Ontology matching is an important tool for the integration of related knowledge sources and the use of available background knowledge is a powerful addition to the unassisted mapping process. In this paper we have shown that for the field of anatomy UMLS, FMA-OBO and WordNet are reliable resources, particularly when combined. The specific strengths of each cover the others' weaknesses, as is elaborated upon in the discussion.

The next step is to start using these external reference sources that we have evaluated. They will provide a basis for the actual mapping of the two species used in this test case. From there we plan to expand our collection of mapped species ontologies, with human being the obvious next addition.

References

[1] http://www.informatics.jax.org/

[2] http://zfin.org/cgi-bin/webdriver?MIval=aa-ZDB_home.apg

[3] http://flybase.org/

[4] http://www.sofg.org/sael/

[5] http://precedings.nature.com/documents/3546/version/1

[6] Ghazvinian A, Noy NF, Musen MA. Creating Mappings for Ontologies in Biomedicine: Simple Methods Work. *AMIA Annu Symp Proc* 2009

[7] Aleksovski Z, Ten Kate W, Van Harmelen F. Exploiting the structure of background knowledge used in ontology matching. *Proc. of the Intl. Workshop on Ontology Matching*: 13-24, 2006

[8] Jean-Mary YR, Shironoshita EP, Kabuka MR. Ontology Matching with Semantic Verification. *Web Sem: Science, Services and Agents on the WWW* 7(3):235-251 2009

[9] Marquet G, Mosser J, Burgun A. A method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies: The case of OBO disease ontologies. *Int J of Med Inf* 76(S3): S353-S361 2007

[10] Bodenreider O, Hayamizu TF, Ringwald M, De Coronado S, Zhang S. Of Mice and Men: Aligning Mouse and Human Anatomies. *AMIA Annu Symp Proc* p.61-65 2005

[11] Bodenreider O, The unified medical language system (UMLS): integrating biomedical terminology. *Nucl Acids Res* 32: D267-D270, 2004

[12] http://sig.biostr.washington.edu/projects/fm/index.html

[13] http://www.obofoundry.org/cgi-bin/detail.cgi?id=fma_lite

[14] George A. Miller. WordNet: A Lexical Database for English. *Comm. of the ACM* 38(11):39-41, 1995

[15] Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol* 6(3):R29, 2005

[16] http://zfin.org/zf_info/anatomy/dict/sum.html

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

[18] Giles, J. Internet encyclopaedias go head to head. *Nature* 438(7070):900-1, 2005

[19] Martini FH, Timmons MJ, Welch K. Human Anatomy Int. Ed. 5th edition (2006)

[20] Nolte J. The Human Brain: An Introduction to its Functional Anatomy. 5th edition (2002)