Issues in learning an ontology from text

Christopher Brewster, Simon Jupp§, Joanne Luciano¶, David Shotton# Robert Stevens§§, and Ziqi Zhang*

**University of Sheffield, §University of Manchester, ¶Harvard University, #University of Oxford*

ABSTRACT

Ontology construction for any domain is a labour intensive and complex process. Any methodology that can reduce the cost and increase efficiency has the potential to make a major impact in the life sciences. This paper describes an experiment in ontology construction from text for the Animal Behaviour domain. Our objective was to see how much could be done in a simple and rapid manner using a corpus of journal papers. We used a sequence of text processing steps, and describe the different choices made to clean the input, to derive a set of terms and to structure those terms in a hierarchy. We were able in a very short space of time to construct a 17000 term ontology with a high percentage of suitable terms. We describe some of the challenges, especially that of focusing the ontology appropriately given a starting point of a heterogeneous corpus.

1 INTRODUCTION

Ontology construction and maintenance are both labour intensive tasks. They present major challenges for any user community seeking to use sophisticated knowledge management tools. One traditional perspective is that once the ontology is built the task is complete, so users of ontologies should not baulk at the undertaking. The reality of ontology development is significantly different. For some large, widely used ontologies, such as the Gene Ontology (Ashburner et al. 2000), a manual approach is effective even if very expensive. For small, scientific communities with limited resources such manual approaches are unrealistic. This problem is all the more acute as research in many areas, including the life sciences, is moving to an e-science industrialised paradigm.

The work presented in this paper concerns the semi-automatic construction of an ontology for the *animal behaviour* domain. The animal behaviour community has recognised the need for an ontology in order to annotate a number of data sets. These data sets include texts, image and video collections. In a series of workshops[1], an initial effort has been made to construct an ontology for the purposes of applying annotations to these data sets. The current Animal Behaviour Ontology (ABO) has 339 classes and the top level structure is shown in Figure 1.

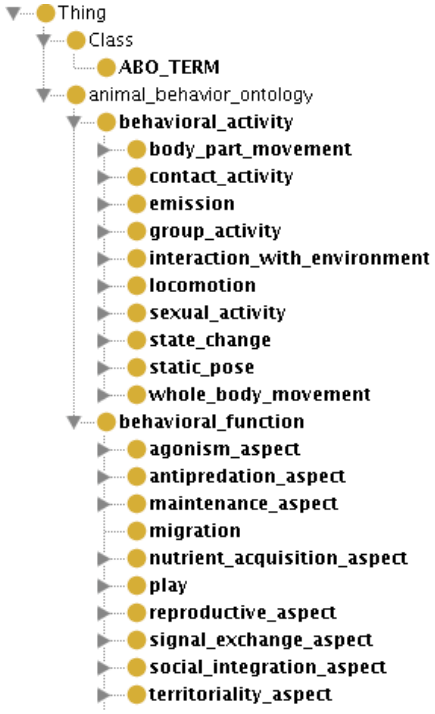

Figure 1 Top level terms in the Animal Behaviour Ontology

While considerable effort has already gone into the construction of the Animal Behaviour Ontology, its limited size raises the important question as to whether it is more appropriate to slowly build an ontology entirely by hand, and have its potential expansion led by user demand, or whether to rapidly build a much larger ontology based on the application of a variety of text processing methods, and tidy or clean the output. With community engagement comes growth, but there is a question of stimulating engagement through some critical mass of useful ontology. The former approach is the standard approach and has been used successfully in cases such as the Gene Ontology, but becomes more challenging as the size and complexity of the ontology increases. On the other hand, while much has been written about automatic ontology learning, most such work has been undertaken in non-biological domains, or in rather abstract contexts (Cimiano et al. 2005; Brewster et al. 2007; Navigli and Velardi 2004). Although such research is called "ontology learning" in reality, given the limitations of Natural Language Processing, the outputs have been structured

* To whom correspondence should be addressed.

[1] For further details cf. http://ethodata.comm.nsdl.org/

Language Processing, the outputs have been structured vocabularies organised in taxonomic hierarchies. This might be considered a major defect if it were not that a) most ontologies are used for labelling/annotation purposes rather than for computational inference, and b) a hierarchically structured vocabulary based on the actual terminology used by a community is a major step towards the creation of a formal ontology. Thus in our view, the construction of formal ontologies of the type needed for driving semantic applications should be considered to involve a significant manual step following the automated process (Luciano and Stevens 2007; Stevens et al. 2007).

In the research reported here, we chose to see how far we could go in the context of limited resources. We approached the challenge as being one to construct a controlled or structured vocabulary as quickly as possible, with minimal effort, and then allow subsequent efforts to clean up the output of this exercise. At one level, we have tried to assess how much effort is worth investing and what is the balance of cost and benefit. A greater understanding of what is the best and most effective methods will in the longer term not only facilitate the creation of useful ontologies for scientific domains with limited resources, but will also facilitate the growing issue of maintenance and upkeep of ontologies as a whole.

2 METHODOLOGY

2.1 The Data Set

It has been argued elsewhere that the only effective way to build representative ontologies for a given domain is through the use of text corpora (Brewster, Ciravegna, and Wilks 2001), and in our case we were able to have access to a considerable corpus of journal articles from the journal *Animal Behaviour*, published by Elsevier. This consisted of articles from Vol 71 (2006) to Vol 74 (2007), containing 623 separate articles. We were given access to text, PDF and XML versions together with a corresponding DTD. We used the XML version for the procedures which are described below.

2.2 From text to ontology

1. Clean text was extracted from the XML files. Using the information from the structured markup, we excluded all author names, affiliations and addresses, acknowledgements, and all bibliographic information, except for the titles of the cited papers.

2. A number of stop word lists and gazetteers were used to further remove noise from the data. We excluded person names as noted above and also through the use of a gazetteer, animal names based on a short list derived from the LDOCE[2], and place names using another gazetteer.

3. A lemmatizer was used to increase coverage (Zhou, Xiaodan Zhang, and Hu 2007). In some cases this generated some noise due to imperfections in the lemmatizer but overall it reduced data sparsity.

4. Five different term extraction algorithms were applied as described in (Ziqi Zhang et al. 2008). The chosen term recognition algorithms were ones that selected both single and multi-word terms as we believe that desirable technical terms are of both sorts. The algorithms were applied to each subsection of the journal article as well as to the whole. This allowed us to look at the terms from different sections of the articles (abstract, introduction, materials and methods, conclusion, etc.). as we aimed to build an ontology of animal behaviour, the terms found exclusively in the "Materials and Methods" section were removed from further consideration. Such terms are the subject of a different ontology.

5a. We then used a set of regular expressions to filter the candidate terms. A regular expression was constructed that looked for terms that ended in *behaviour, display, construction, inspection*, etc. It also included some very generic regular expressions looking for terms that ended in *-ing* and *–ism*. The regular expression used for term selection is available on the website accompanying this research[3].

5b. The step described in 5a. involved quite specific domain knowledge. To have an alternative procedure that does not involve any domain knowledge, we used a voting algorithm to rank the terms and weight them for distribution across the corpus. This was calculated by taking the mean rank for each term and multiplying by the document frequency. From the resulting rankings terms were selected for the subsequent steps (to parallel those extracted by the regular expression).

6a. There are a number of methods that can take a set of terms and try to identify ontological (taxonomic) relations between the terms (Cimiano, Pivk, Schmidt-Thieme, and Staab 2005; Brewster 2007). Most methods suffer from low recall. So in our approach we chose to use the method used in the literature with highest recall – string inclusion. This means that a term A B ***IS_A*** B, and A B C ***IS_A*** (B C and A C) ***IS_A*** C. The resulting ontology was saved in the Web Ontology Language (OWL).

6b. The same method as 6a. was applied to the output of 5b.

7a. and 7b. The resultant ontologies were then filtered for their top level terms i.e. children of THING. A technique used extensively in the ontology learning community is that of using lexico-syntactic patterns (or Hearst patterns (Hearst 1992)) to either learn or test for a candidate ontological relation (Brewster et al. 2007). In this case, we tested each top

[2] The Longman Dictionary of Contemporary English. Our thanks to Louise Guthrie for providing this.

[3] http://nlp.shef.ac.uk/abraxas/animalbehaviour.html

level term in each ontology as to whether it was a kind of *behaviour, activity* or *action* using the Internet as an external resource. Thus we constructed phrases such as the following: *"behaviours such as biting"* (found) or *"behaviours such as dimorphism"* (not found).

3 RESULTS

Figure 2 Partial subtree from ontology at Step 6a.

A total of 64,000 terms were extracted from the whole corpus of 2.2 million words. From this the regular expression extracted 10,335 terms. These included animal behaviour terms, but also included non-animal behaviour terms. The regular expression was designed to capture a large number of terms such as *begging, foraging, dancing, grooming, burrowing, mating*. Due to its crudity it also picked up non-behavioural terms with similar endings: *-bunting, -herring, dichromatism, dimorphism*.

The ontology produced by Step 6a. resulted in an artefact of 17776 classes, of which 1295 classes are top level (i.e. direct children of `OWL:THING`). The ontology produced by Step 6b. from the 10,335 terms selected by the voting algorithm in step 5b. resulted in an artefact of 13,058 classes, of which 2535 classes were top level. The ontologies mentioned here are available on the web site accompanying this paper[4]. A screen shot of the sub tree concerning *call* from ontology 6a. is shown in Figure 2.

The filtering process described in Step 7a. resulted in 383 top level terms being removed leaving 912 immediate descendants of `OWL:THING`. Top level classes that were filtered out by this method included terms such as *stocking referencing, holding, attraction, time, schooling, movement, pacing, defending, smashing, loading, matricide.* The parallel process in 7b. resulted in 649 top level classes being removed, leaving 1886.

A sample of the terms excluded by step 5a. has been evaluated by a biologist (Shotton). Of the 56,000 terms excluded, a random sample of 3140 terms were manually inspected. Of these 7 verbs and 42 nouns were identified as putative animal behaviour-related terms. These included terms such as *forage, strike, secretion, ejaculate, higher frequency yodel, female purring sound,* etc. The low number of significant excluded terms shows that our approach has a *Negative Predictive Value* of 0.98, and a *Recall* of 0.905. We have yet to determine the precision of this approach due to the need for large scale human evaluation of the selected terms.

4 DISCUSSION

A key challenge in the process of learning an ontology from texts is to identify the base units, i.e. the set of terms which will be used as labels in the ontology's class hierarchy. This problem has been largely ignored in the NLP ontology learning literature. The problem of constructing an ontology from a data set such as the one we were using is that in effect there are a number of different domain ontologies represented in the text. In the case of our corpus from the journal *Animal Behaviour*, there existed terms reflecting *experimental methods, animal names, other named entities (places, organisations, people),* etc in addition to behaviours. Such domains are obviously pertinent to animal behaviour (there are species specific behaviours), but the terms exclusively from these domains belong to separate ontologies. The linking together of these separate domains within one ontology is a further step in the process of ontology building.

In order to construct an ontology of animal behaviour from such a heterogeneous data set, one must focus the term selection as much as possible. In order to do this we used first a manually constructed set of regular expressions, an approach which is dependant on domain expertise. As an alternative, for the sake of comparison, we selected the same number of terms using the term recognition voting approach. The ontology generated by this latter approach re-

[4] http://nlp.shef.ac.uk/abraxas/animalbehaviour.html

sulted in less complexity because it included fewer multi-word terms, which using our string inclusion method had generated further intermediate concepts and a richer hierarchy when using the terms identified by regular expressions.

Our initial evaluation of the terms excluded by the regular expressions shows that very few of the omitted terms were significant from an expert's perspective. Our approach will tend to high recall and low precision so there are certainly a significant number of terms included that would need subsequent manual exclusion. A brief consideration of Figure 2 shows a number of terms that would need to be excluded: *g_call, lower call*, etc.

Nevertheless, the resulting ontologies, especially after filtering the top level terms, contains a large number of useful taxonomic fragments even if there is quite a lot of noise. Part of the principle of our approach, as noted in the Introduction, is that it is far easier to collect a large set of potentially significant ontological concepts automatically and then eliminate the noise than to slowly build up a perfectly formed but incomplete set of concepts but which inevitably will exclude a lot of important domain concepts. Such an artefact is far from a formal ontology but is nonetheless useful as a step towards a taxonomic hierarchy for the annotation of research objects, and as a stepping-stone to a more formal ontology. While we still have to undertake a full evaluation, initial assessments indicate the ontologies derived using the regular expressions are cleaner and of greater utility.

The limitations of our approach may be summarised as follows: a) there is a certain amount of noise in the resulting ontologies (which we specify more precisely in future work), b) some effort is involved in *focussing* the ontology produced (i.e. to exclude terms that properly belong to another domain/ontology), c) the result is only taxonomic – the use of string inclusion implies an ISA hierarchy although careful inspection shows that this is not always the case.

The significance of our approach is that it is very quick and easy to undertake. The results produced are very useful, both in themselves as a knowledge discovery exercise in a scientific domain, and as a stepping stone to a more rigorous or formal ontology. The very low effort involved in the process means that this type of data collection could be used in all cases when building ontologies from scratch. We also propose this approach as being a significant tool in ensuring ontologies are up to date and are current with the terminology of a domain.

Future work will include applying the full Abraxas methodology (Brewster et al. 2007) to construct the richest possible structure from the existing ontology. We plan a more extensive evaluation of the noise present i.e. terms that should be excluded. At a more fundamental level, we need to consider how appropriate it is to use terms derived from a corpus for the building of an ontology in contrast to a formally and rigorously hand built ontology.

ACKNOWLEDGEMENTS

We would like to thank Anita de Ward of Elsevier for making the text available from the Journal *Animal Behaviour*. This work was supported by the AHRC and EPSRC funded Archeotools project (Zhang), the Companions project (www.companions-project.org) (IST-FP6-034434) (Brewster), and the Sealife project (IST-2006-027269) (Jupp).

REFERENCES

Ashburner, M. et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, no. 1:25-29.

Brewster, Christopher. 2007. Mind the Gap: Bridging from Text to Ontological Knowledge. Department of Computer Science, University of Sheffield.

Brewster, Christopher, Fabio Ciravegna, and Yorick Wilks. 2001. Knowledge Acquisition for Knowledge Management: Position Paper. In Proceeding of the IJCAI-2001 Workshop on Ontology Learning, Seattle, WA http://www.dcs.shef.ac.uk/~kiffer/papers/ontolearning.pdf.

Brewster, Christopher et al. 2007. Dynamic Iterative Ontology Learning. In Recent Advances in Natural Language Processing (RANLP 07), Borovets, Bulgaria.

Cimiano, Philipp, Aleksander Pivk, Lars Schmidt-Thieme, and Steffen Staab. 2005. Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence, IOS Press http://www.aifb.uni-karlsruhe.de/WBS/pci/OLP_Book_Cimiano.pdf.

Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING 92), Nantes, France, July 1992.

Luciano, Joanne S, and Robert D Stevens. 2007. e-Science and biological pathway semantics. BMC Bioinformatics 8 Suppl 3:S3.

Navigli, Roberto, and Paula Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Websites. Computational Linguistics 30, no. 2:151-179.

Stevens, Robert et al. 2007. Using OWL to model biological knowledge. Int. J. Hum.-Comput. Stud. 65, no. 7:583-594.

Zhang, Ziqi, Jose Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco.

Zhou, Xiaohua, Xiaodan Zhang, and Xiaohua Hu. 2007. Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI) http://www.dragontoolkit.org/dragontoolkit.pdf.