Towards Finding Animal Replacement Methods

Nadine Dulisch and Brigitte Mathiak(✉)

GESIS - Leibniz Institute for the Social Sciences,
Unter Sachsenhausen 6-8, Cologne, Germany
{nadine.dulisch,brigitte.mathiak}@gesis.org

Abstract. Protecting animal rights and reducing animal suffering in experimentation is a globally recognized goal in science. Yet numbers have been rising, especially in basic research. While most scientists agree that they would prefer to use less invasive methods, studies have shown that current information systems are not equipped to support the search for alternative methods. In this paper, we outline our investigations into the problem. We look into supervised and semi-supervised methods and outline ways to remedy the problem. We learned that machine assisted methods can identify the documents in question, but they are not perfect yet and in particular the question about gathering sufficient training data is unsolved.

Keywords: Classification · Animal welfare

1 Introduction

Researchers from Life Sciences that contemplate to use animal testing are motivated by ethical, financial and often legal incentives to try and find alternatives that enable them to answer the same research questions, but with less animal involvement.

Unfortunately, current strategies for literature search do not support search for animal test alternatives very well. Dutch animal welfare officers have been asked for their strategies in finding alternative methods [2]. None found this task easy and reported that the most successful way of finding good alternatives was word of mouth.

When interviewing experts in finding such documents, it becomes clear that there are two criteria. *Similarity* is how close the document is to the experiment we want to replace. *Relevance* measures how likely it is that the document describes a method that causes less animal suffering. In order to give a comprehensive list of candidate documents, we need to take both criteria into account.

2 Related Work

Our attempt at solving the animal test replacement problem is not the first one. Go3R[1] is a semantic search engine based in PubMed and ToxNet[2] prioritizing

[1] http://www.gopubmed.org/web/go3r/.
[2] http://toxnet.nlm.nih.gov.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 595–598, 2017.
DOI: 10.1007/978-3-319-67008-9_51

3R and toxicology. While the toxicology use case is interesting and useful, the focus on recall makes it hard to handle basic research questions, which typically do not fit semantic categories as neatly. AltBib[3] is not an independent search engine, but rather suggests query term expansions that, among other factors, utilize the MeSH classification for animal testing alternatives. This classification is not systematically used to tag all methods that are developed as alternatives, but seems to focus on documents about animal testing alternatives on a meta level. In 2015 there were less than 3000 documents with that MeSH term.

Table 1. Overview over # of positive, negative and not classified instances for the different use cases.

Use case	Reference document (PMID)	Relevance			Similarity			Animal test		
		+	-	/	+	-	/	+	-	/
1	16192371	13	85	2	15	77	8	13	70	17
2	11932745	21	70	9	13	78	9	47	39	14
3	11489449	13	80	7	4	75	21	37	55	8

3 Corpus

To our knowledge, there is no corpus available to use as training data, a problem that has been plaguing Information Retrieval from the very beginning [1]. We structured our corpus around individual use cases, mimicking the application process for getting a permission to conduct a specific animal experiment. The starting point was a document describing an animal test (reference document), which was chosen by the domain expert. To search for possible alternative methods, we used the PubMed functionality to find similar documents based on substring similarity[4].

The first 100 hits of documents similar to the reference document were downloaded and then assessed by the domain expert according to criteria the expert set down beforehand. Additionally to the aforementioned dimensions of result similarity and replacement relevance, we also asked the expert to give us information on whether the document described animal experiments or not.

Non-classification occurred when there was not enough information to make a sensible decision, e.g. missing abstract, missing relevant information, non-available full text or, in rare cases, when the expert felt not knowledgeable enough about the domain to make a judgement call. In the following experiments, we only used the classified documents.

[3] http://toxnet.nlm.nih.gov/altbib.html.
[4] http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Computation_of_Similar_Articl.

For each document we collected the following metadata: title, abstract, PubMed Central ID, URLs, journal name, availability status of the full text and MeSH term information.

Table 1 gives an overview over our created use cases. The table shows the number of positive (+) and negative (−) instances included in the datasets. The datasets include only few positive instances, leading to strong bias in learning.

4 Experiments

4.1 Classification

What we are most interested in, is if trained algorithms are able to distinguish between positive and negative instances. In this experiment, we compared the prediction performance of standard data mining algorithms, which included J48 (C4.5 decision tree), JRIP (Propositional rule learner), SMO (Sequential minimal optimization), Naïve Bayes, Bayes Net and LWL(Locally weighted learning). We conducted the experiments applying the data mining software Weka[5] (version 3.6) and used Weka's implementation of the aforementioned algorithms. For classification we used Weka's "FilteredClassifier", applying the "StringToWordVector" filter to handle string attributes. This filter transforms string attributes into an attribute set that represents word occurrence information[6]. We used leave-one-out evaluation for all experiments, based on the original dataset. For the results see Table 2.

Table 2. Average F-Score over all three use cases, differentiated after algorithm and metric. Note that the F-Score is calculated from the point of view of the positive instances, therefore the expected value for random choice is very low due to the bias.

Target attribute	Unbalanced dataset					
	J48	JRIP	Naïve Bayes	Bayes Net	SMO	LWL
Relevance	0.51	0.36	0.44	0.66	0.52	0.42
Similarity	0.14	0.08	0.36	0.28	0.18	0.07
Animal test	0.79	0.87	0.91	0.87	0.94	0.83

We immediately discovered that the unbalancedness of the datasets, with as little as 4 positive examples were creating serious problems as especially relevant and similar documents were only rarely correctly classified (cf. Table 2). This is particularly devastating for similarity, where most results are worse than or close to random. Results for relevance are not ideal, but clearly better than random. Animal tests can be detected quite reliably, but positive and negative instances are much better distributed, as you can see in Table 1.

[5] http://www.cs.waikato.ac.nz/ml/weka/.

[6] http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/StringToWordVector.html.

Table 3. F-Score for semi-supervised learning. Averaged over all use cases. Number in parentheses is the original value.

Target attribute	Naïve Bayes	SMO
Relevance	0.56 (0.44)	0.53 (0.52)
Similarity	0.38 (0.36)	0.43 (0.18)

Semi-supervised Learning. As discussed before, we had only comparably few labeled documents available, and in real life, we might have even less. What we have not leveraged so far is that we had a high number of unlabeled documents available to us that fit the general topic. Following the self-training methodology laid out by [3] we used a semi-supervised learning approach in which unlabeled documents were used to counteract the scarcity of training data.

The semi-supervised approach improves results for relevance compared to the original values for the unbalanced dataset. As Table 3 shows, the F-Score value for relevance increases for both top algorithms, but only moderately. The SMO F-Score for similarity, however, raises more significantly, which seems to indicate that the lack of training data impacted the classifiers ability to successfully predict similarity.

5 Conclusions and Future Work

While we do not have a workable prototype for up-ranking animal replacement methods yet, we believe we have made important inroads and identified some roadblocks. On the bright side, we have shown that relevant documents can be found with machine learning given enough training data. Methods to reduce the need for training data have been tested and were found to be successful.

A more direct approach would be to use un-supervised methods of finding similar documents. Bibliometric methods seem hopeful, but positive and negative effects overlap and cancel each other out. We tried using classifiers across use cases, but without any improvements. On all fronts, it becomes clear that more training data is needed.

References

1. Jones, K.S., van Rijsbergen, C.J.: Information retrieval test collections. J. Documentation **32**(1), 59–75 (1976)
2. van Luijk, J., Cuijpers, Y., van der Vaart, L., de Roo, T.C., Leenaars, M., Ritskes-Hoitinga, M.: Assessing the application of the 3rs: a survey among animal welfare officers in The Netherlands. Lab. Anim. **47**(3), 210–219 (2013)
3. Zhu, X.: Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison (2005)