Classification of Study Region in Environmental Science Abstracts

Jared Willett,[♠] **Timothy Baldwin,**[♠♡] **David Martinez**[♡] and **Angus Webb**[♢]
♠ Department of Computing and Information Systems
♡ NICTA Victoria Research Laboratory
♢ Department of Resource Management and Geography
The University of Melbourne, VIC 3010, Australia
jwillett@student.unimelb.edu.au, tb@ldwin.net,
davidm@csse.unimelb.edu.au, angus.webb@unimelb.edu.au

Abstract

One of the potentially most relevant pieces of metadata for filtering studies in environmental science is the geographic region in which the study took place (the "study region"). In this paper, we apply support vector machines to the automatic classification of study region in a dataset of titles and abstracts from environmental science literature, using features including frequency distributions of resolved toponyms and a bag of word unigrams. We found that we can determine the study region with high accuracy, with the strongest classifier achieving an accuracy of 0.892 combining toponym resolution from DBpedia and GeoNames with the bag-of-toponyms features.

1 Introduction

One of the potentially most relevant pieces of metadata for filtering studies in environmental science is the region in which the study took place, as users making queries are often looking for studies performed in a specific area. However, bibliographic databases do not systematically include information on study location. The Eco Evidence database, a compendium of literature citations and linked evidence items that is used for evidence synthesis in companion software (Webb et al., 2011), is one such system for which location information is very helpful. However, the manual annotation of such metadata over large quantities of literature is a tedious and time-consuming task.

One possible solution to this issue is to have this information automatically extracted with the aid of natural language processing (NLP) techniques. The abstracts of studies, which are commonly available in bibliographic databases, frequently contain geographic references of various granularities. If identified and resolved, these toponyms provide the potential to make a strong estimation of the overall location of the study. In these experiments, we evaluate the performance of various NLP techniques for automatic classification of the study region in environmental science literature abstracts.

Beyond the aim of being able to quickly assemble a collection of literature from a given area, our motivation in applying NLP to automatically extract information from environmental science literature is driven by our interest in moving towards an evidence-based model of decision-making in the environmental sciences (Sutherland et al., 2004). Similar to evidence-based medicine (Sackett et al., 1996), such a model relies heavily on systematic literature reviews as a means of synthesizing evidence from the literature. The Eco Evidence database (Webb et al., in press) is a compendium of literature citations and linked evidence items that is used for systematic review and evidence synthesis in companion software (Webb et al., 2011). The database is in active use in a number of research projects currently, and evidence therein has also formed the basis of several published systematic reviews (Webb et al., 2012). However, all evidence in the database is currently manually annotated.

2 Background Work

Our motivation in applying NLP to automatically extract information from environmental science literature is driven by our interest in moving towards an evidence-based model of decision-making in the environmental sciences (Sutherland et al., 2004), similar to evidence-based medicine (Sackett et al., 1996). Our work is directly motivated by the possibility of streamlining the population of the Eco Evidence database by auto-

Jared Willett, Timothy Baldwin, David Martinez and Angus Webb. 2012. Classification of Study Region in Environmental Science Abstracts. In *Proceedings of Australasian Language Technology Association Workshop*, pages 118−122.

matically extracting location information, but has wider potential application to other bibliographic databases where there is a geospatial dimension to the data.

Comparable work in the biomedical domain has focused on the automatic extraction of Medical Subject Headings (MeSH) terms in abstracts (Gaudinat and Boyer, 2002), labeling documents based on specific terms in the abstract which are to be resolved to more general categories.

The unique opportunities and challenges specific to retrieving geospatial information have been well documented, particularly in the context of geospatial information retrieval where queries and documents have a geospatial dimension (Santos and Chaves, 2006). Aside from finding locations in the text, the disambiguation of what exact location a term in a text is referring to presents a unique challenge in itself, and a variety of approaches have been suggested and demonstrated for this task (Overell and Rüger, 2006).

The methodology described in this work is based on a standard approach to geographic informaton retrieval, which was demonstrated by Stokes et al. (2008) in their study of the performance of individual components of a geographic IR system. In particular, the named entity recognition and toponym resolution (TR) components are the basis for all the main classifiers in this study.

3 Method

3.1 Dataset

The dataset for these experiments consists of the titles and abstracts for 4158 environmental science studies recorded in the Eco Evidence database. One such sample abstract (Fu et al., 2004) can be read below:

> Title: Hydro-climatic trends of the Yellow River basin for the last 50 years
>
> Abstract: Kendall's test was used to analyze the hydro-climatic trends of the Yellow River over the last half century. The results show that: ...[1]

Study regions for these papers have been manually annotated, providing a gold standard for purposes of training and evaluation. The study region can be chosen from ten different options: Europe, Australia, Africa, Antarctica, Asia, North America, South America, Oceania, Multiple and Other. The dataset is not evenly distributed: North America is the most commonly annotated study region, covering 41.5% of the studies, while other classes such as Antarctica and Other were extreme minorities. Oceania represents all countries contained in Australasia, Melanesia, Micronesia and Polynesia, with the exclusion of Australia (which, as a continent, has its own category). 'Multiple' refers to studies done across multiple regions, and 'Other' is used for studies where no particular region is evident or relevant to a work (i.e. a literature review). These two labels present difficulty for methods based on toponym resolution, as studies with toponyms from multiple regions or none at all are often still considered to be located in one continent. However, Multiple and Other are minority labels, comprising only 3.5% and 0.2% of the dataset respectively.

3.2 Named Entity Recognition

The first component of our system involves extracting references to locations contained in the abstract, a task which we approach using named entity recognition (NER). NER is an NLP task in which we seek to automatically extract 'named entities', which refer to any term in a body of text that represents the name of a thing considered an instance of one of a predefined set of categories.

Our first experiments focused on evaluating the performance of the off-the-shelf 3-class model of the Stanford NER system (Finkel et al., 2005) in detecting relevant named entities in the titles and abstracts. The NER system classifies identified entities as people, locations or organizations. For our task, only named entities that are locations are relevant, thus only these entities are extracted and evaluated.

3.3 Toponym Resolution

Once the named entities tagged as locations for each abstract were collected, we experimented with resolving each location to its corresponding continent using two different databases of geospatial entities. Two methods were employed for each database: (1) observing only the top result to resolve the location; and (2) returning the frequency distribution of the top-five results.

In each classifier where tags were resolved to continents, we experimented with using each sys-

[1] The sample abstract has been truncated here, but contains no further toponyms.

tem separately as well as in combination, simply combining together the results from the two databases.

3.3.1 DBpedia

First, we resolve toponyms with DBpedia (`http://www.dbpedia.org`), a database of structured content extracted from Wikipedia. If the page retrieved for a given toponym has geographic coordinates available, these are extracted and checked against a set of non-overlapping bounding boxes, which were manually constructed by setting one or more ranges of longitude and latitude for each possible label except 'Multiple' and 'Other'. If the extracted coordinates are within the range of one of the bounding boxes, the corresponding label is applied to the term.

For terms with multiple meanings, DBpedia will contain a disambiguation page. For the top-result TR approach, in the event that coordinates are unavailable for the first possibility on the disambiguation page, no resolution is recorded for the term. For the top-5 approach, we continue to look for results until all disambiguations have been exhausted or five resolutions have been found.

3.3.2 GeoNames

Second, we resolve toponyms with GeoNames (`http://www.geonames.org`), a gazetteer which collects data from a wide variety of sources. A query was done for each toponym using the GeoNames search function, which directly provides a ranked list of results with continent codes.

3.4 Majority Vote

As a baseline, we use only the retrieved continents from either DBpedia, GeoNames or both, and determine the final classification by a simple majority vote. When there is a tie in the top number of resolutions, the continent that appears most frequently in the training data is chosen. If the classifier is unable to resolve any toponyms for a given instance, the majority class label in the training data (which is consistently North America, across all folds of cross-validation) is used as a backoff.

3.5 SVM Classification

All our supervised classifiers are based on support vector machines (SVMs), using LibSVM (Chang and Lin, 2011). With SVMs, instances are represented as points in n-dimensional space, with each dimension representing a different feature, and the classification of test instances is done based on which side of a binary dividing hyperplane the instance falls on. In all our experiments, we use a linear kernel, and all other LibSVM parameters are set to the default. The SVM method is adapted to the multi-class task in LibSVM using the "one-against-one" method, in which binary classification is used between each two candidate labels and the label for which the instance is classified to the highest number of times it is selected. In this section, the features used to construct the vectors are described.

Classifier	F-score
Majority class	0.415
Oracle	0.969
Bag-of-Toponyms (BoT)	0.834
Bag-of-Words (BoW)	0.729
BoT + BoW	0.773

Table 1: Accuracy for classifiers w/o toponym resolution.

3.5.1 Continent Resolutions

The continent-level results from DBpedia and/or GeoNames were represented as frequency distributions over the number of results for each continent returned for a given instance. When both DBpedia and GeoNames are used, the counts are accumulated into a single frequency distribution.

3.5.2 Bag of Words Features

We used a bag-of-words model in two forms. The first only considered the toponyms as tokens, creating features of a count for each toponym over the full dataset. The second type applied the standard bag-of-words model over all words found in the abstracts.

3.6 Evaluation

In order to establish an upper bound for the task, the first author manually performed the study region classification task over 290 randomly-sampled abstracts classified. The accuracy for this "oracle" method was 0.969. In all of cases where the manual annotation was incorrect, there was insufficient data in the abstract to reasonably deter-

Classifier	DBp:1R	Geo:1R	D+G:1R	DBp:MR	Geo:MR	D+G:MR
Majority Vote	0.802	0.830	0.875	0.788	0.822	0.851
SVM	0.829	0.832	0.877	0.813	0.843	0.862
SVM + BoT	0.879	0.877	0.892	0.873	0.879	0.887
SVM + BoW	0.855	0.862	0.889	0.846	0.868	0.884
SVM + BoT + BoW	0.862	0.868	0.891	0.854	0.873	0.886

Table 2: Accuracy for DBpedia/GeoNames classifiers ("1R" = top-1 toponym resolution; "MR" = multiple resolutions)

mine the location of the study.[2]

Our primary evaluation metric for the overall classification task is classification accuracy. For all classifiers except the oracle annotation, the final scores are the result of 10-fold stratified cross-validation over the dataset.

4 Results

First, we evaluated the token-level accuracy of the NER system over our dataset, to determine its performance in the domain of environmental science. 30 abstracts were selected randomly, and all named entity locations were manually identified. Based on these annotations, the off-the-shelf results for the Stanford NER were a respectable 0.875 precision, 0.778 recall, and 0.824 F-score. One of the most common causes of false positive was species names. The knock-on effect of incorrect or missed tags should be considered as one source of error in the overall classification task.

Table 2 shows the results of the classifiers featuring toponym resolution. Overall, the DBpedia and GeoNames classifiers performed at a similar level, with most GeoNames classifiers slightly outperforming their DBpedia counterparts. When the resolutions from DBpedia and GeoNames were combined, accuracy increased for all classifiers. Combining the results basically doubles the confidence of continents where there is agreement between the databases, which can be particularly helpful given the sparsity of tagged locations for each abstract. Including the top-5 results ("MR" in Table 2) consistently decreased the accuracy, suggesting that noise in incorporating additional possible disambiguations outweighs any gains in capturing ambiguity.

Supervised learning is clearly beneficial to the task, as the majority-vote classifier is consistently outperformed by the SVM classifiers, particularly when bag-of-toponyms and/or bag-of-words features are included. The bag-of-toponyms consistently outperforms the unfiltered bag-of-words, especially when isolated from TR frequency features (shown in Table 2), indicating that including other lexical information provides insufficient additional relevance to outweigh the noise, and that explicitly incorporating geospatial features boosts accuracy. Ultimately, the best-performing classifier utilised the top result from both DBpedia and GeoNames, using the bag-of-toponyms and top-result frequency features, achieving an accuracy of 0.892, well above the accuracy of both the majority class baseline at 0.415 and the simple bag-of-words classifier at 0.729, and only slightly below the human-based upper bound of 0.969. The difference between this best-performing SVM classifer and the majority vote classifier of the same toponym resolution approach was found to be statistically significant ($p = .001$) using randomization tests (Yeh, 2000).

5 Conclusion and Future Work

We have demonstrated that NLP approaches paired with toponym resolution are highly successful at identifying the study region from the abstracts of publications within the environmental science domain, with our best classifier achieving an accuracy of 0.892, compared to a human-based upper bound of 0.969.

Possible future work could include weighting of different toponym granularities, exploiting geo-spatial relationships between identified toponyms, and domain-adapting a NER for the environmental sciences.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Com-

[2] In many cases, the gold-standard annotation was based on the full-text paper, which we do not make use of in this study.

munications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

C.-C. Chang and C.-J. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Ann Arbor, USA.

G. Fu, S. Chen, C. Liu, and D. Shepard. 2004. Hydroclimatic trends of the yellow river basin for the last 50 years. *Climatic Change*, 65(1):149–178.

A. Gaudinat and C. Boyer. 2002. Automatic extraction of MeSH terms from Medline abstracts. In *Workshop on Natural Language Processing in Biomedical Applications*, pages 53–57, Nicosia, Cyprus.

S.E. Overell and S. Rüger. 2006. Identifying and grounding descriptions of places. In *3rd Workshop on Geographic Information Retrieval*, Seattle, USA. SIGIR.

D.L. Sackett, W. Rosenberg, JA Gray, R.B. Haynes, and W.S. Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

D. Santos and M.S. Chaves. 2006. The place of place in geographical IR. In *3rd Workshop on Geographic Information Retrieval*, Seattle, USA. SIGIR.

N. Stokes, Y. Li, A. Moffat, and J. Rong. 2008. An empirical study of the effects of nlp components on geographic ir performance. *International Journal of Geographical Information Science*, 22(3):247–264.

W.J. Sutherland, A.S. Pullin, P.M. Dolman, and T.M. Knight. 2004. The need for evidence-based conservation. *Trends in Ecology & Evolution*, 19(6):305–308.

J.A. Webb, S.R. Wealands, P. Lea, S.J. Nichols, S.C. de Little, M.J. Stewardson, R.H. Norris, F. Chan, D. Marinova, and R.S. Anderssen. 2011. Eco Evidence: using the scientific literature to inform evidence-based decision making in environmental management. In *MODSIM2011 International Congress on Modelling and Simulation*, pages 2472–2478, Perth, Australia.

J.A. Webb, E.M. Wallis, and M.J. Stewardson. 2012. A systematic review of published evidence linking wetland plants to water regime components. *Aquatic Botany*.

J.A. Webb, S.C. de Little, K.A. Miller, and M.J. Stewardson. in press. Eco evidence database: a distributed modelling resource for systematic literature analysis in environmental science and management. In R. Seppelt, A. A. Voinov, S. Lange, and D. Bankamp, editors, *2012 International Congress on Environmental Modelling and Software*, Leipzig, Germany. International Environmental Modelling and Software Society.

A. Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proc. of the 18th International Conference on Computational Linguistics (COLING 2000)*, volume 2, pages 947–953, Saarbrücken, Germany. Association for Computational Linguistics.