Jacquelynn K. Sherman. Automatic metadata generation: a comparison of two annotators. A Master's Paper for the M.S. in L.S degree. April, 2010. 28 pages. Advisor: Jane Greenberg

There is a growing need to develop effective techniques and tools for automatic metadata generation. The research presented in this master's paper compares the annotation functions of NCBO BioPortal with those of HIVE in order to determine whether basic term matching techniques or machine learning techniques produce higher quality results. The research was conducted by selecting a document set and testing it on both annotators. The metadata generated by the annotators was then assessed by three human evaluators in terms of relevance, precision, specificity, and exhaustivity. The research found that on average the results produced by the HIVE annotator, which employed machine learning techniques, had 10 percent higher specificity, 17 percent higher exhaustivity, and 19.4 percent higher precision than the results produced by BioPortal. The paper concludes that the machine learning underlying HIVE produces higher quality results than basic term matching techniques, and that this approach deserves greater research attention.

Headings:

Metadata

Metadata Generation Tools

Metadata - Automatic Metadata Extraction

Metadata Quality

Automatic indexing

AUTOMATIC METADATA GENERATION: A COMPARISON OF TWO ANNOTATORS

by
Jacquelynn K. Sherman

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2010

Approved by

_____

Jane Greenberg

TABLE OF CONTENTS

INTRODUCTION

Automated indexing, metadata generation, and annotation are growing areas of research in library and information science. This is largely due to the explosion of digital resources to be described as well as the high cost, time-ineffectiveness, and inconsistency which accompanies the performance of these tasks manually. The need for development of techniques and tools which can effectively and efficiently generate high quality indexing and annotation is steadily increasing in importance. This is necessary not only for the benefit of institutions, repositories, and others generating this data for information resources; it also benefits users of information by improving access, retrieval, and availability of resources for the user.

The focus of this study is automated subject metadata generation. More specifically, this research will focus on two methods of automated annotation, basic term matching and machine learning, in an attempt to discover which method produces the highest quality results. The hypothesis of this study is that automated annotation which employs machine learning will generate results of higher quality than those produced by automatic annotation based on term matching.

As pointed out by Miksa (1998), the modern library movement was precipitated by a paradigm shift in the focus of libraries from education to information accessibility.

It was this shift which eventually led to the rise of computer science in information retrieval. Opportunities presented by new technology allowed for the creation of came to be known in the early 1990s as "digital libraries" (Marchionini, 1995). Marchionini notes that for those in the library science field, these digital libraries made necessary exploration of new ideas regarding classification and an increased importance of electronic means for managing resources. These developments in technology and library science have contributed to the existence of and growing research in automatic indexing as well as other automatic metadata generation.

The information explosion coupled with increased demand for information has heightened the importance of metadata. Metadata is essential for the organization of both physical and digital forms of this ever-growing wealth of information. In addition to the classification and organization of these resources, metadata is also important for effective retrieval; after all, information which is stored but cannot be located or accessed is of little use. Providing access to an array of seemingly infinite information is an increasingly difficult task, however, and requires a great deal of resources. Given that human indexers and metadata professionals are quite costly, and manual metadata generation is time-consuming, it is becoming more and more imperative that tools be created to automatically generate metadata, to be used both in place of human professionals and as tools to assist in the creation process. Perhaps even more important is the need to develop high quality generation applications in order to reduce noise, which Shen (2007) describes as interference in the form of inaccurate or unhelpful records, in order to ensure efficiency when metadata generation is solely automatic as well as enhance usefulness to human professionals. One way to address this need is to compare existing operational

automatic metadata generation systems, in order to determine which methodologies provide the best results, and consider means for improvement. This is the central goal of the research presented in this paper. This paper includes a literature review discussing the demand for automatic indexing and metadata generation; existing research on the topic; and recommendations for the automatic generation of metadata. Following the literature review is a description of the methodology and the annotators used in the study. The paper concludes by highlighting results from the study, discussing the findings, noting limitations, and identifying future research directions.

LITERATURE REVIEW

A survey of the literature regarding automatic metadata generation reveals a growing need for its development in various settings. This section reviews literature that explores the necessity for automatic indexing and metadata generation. Additionally, it discusses the research which currently exists regarding automatic metadata generation as well as recommendations for creating and using metadata, both manually and automatically generated, to meet the needs of content creators, repositories of various types, and users.

Research involving automatic metadata generation spans many disciplines and formats. The case for subject metadata generation is similar to that of automatic indexing and can be broken down into four main factors. The first factor which makes necessary automated metadata generation is the growing number of resources and information produced and handled by institutions and repositories. This ever expanding amount of

information must be cataloged, indexed, and optimized for retrieval to ensure that users can access it.

A prevalent opinion in much of the literature submits that the method of indexing which produces the highest quality results is that performed by a human indexer. As Anderson and Perez-Carballo (2001) point out, the primary aspect of human indexing that makes essential the development of efficient automatic indexing methods is the cost. Identification of the cost of metadata generation of various types as a growing problem is not a new development; in 1964, O'Conner, in discussing the cost of human indexing, noted an estimate claiming that "subject indexing accounts for about three quarters of the cost of operating a retrieval system." This high figure highlights a need to explore automated indexing and subject metadata generation as an alternative and effective means. This growing dilemma has not gone unnoticed in more recent literature; the expense of human indexing has increased over time, while automatic indexing continues to become cheaper and more effective (Anderson & Perez-Carballo, 2001).

Another issue with indexing and metadata performed manually by humans deals with standardization and consistency. The products of two human indexers are unlikely to be the same, and even two separate instances of indexing by a human indexer of the same document can differ (Sampson and Babarczy, 2008). An automatic indexer, however, will produce the same indexing results for the same document consistently. Some methods of machine learning have been shown to improve results produced by automatic indexers in documents with a length equivalent to that of an abstract (Salton, 1970).

The circumstances which involve human indexing, however, further contribute to the need for efficient automatic indexing and metadata generation. Perhaps the most salient of these circumstances is the amount of time taken to index resources manually. The creation of subject metadata for resources is very time-consuming, which proves problematic since it means a greater delay between resource acquisition and the use of the acquired resource (Sampson & Babarczy, 2008).

Indeed, the issues associated with manual indexing and metadata generation done by humans seem to point to development of automatic methods as essential to resolving this "metadata bottleneck" as Liddy, et al (2002) has described it; and while indexing and metadata generation done manually by humans has often been presented in the literature as having a much higher quality than that produced automatically, this perspective is not unanimous. Some research has indicated that the disparity in quality between human generated and automatically generated metadata is not as great as other would make it seem; Cardinaels (2005), for example, concludes in his exploration of automatic web indexing services that metadata can be generated automatically in certain situations "without a great loss of accuracy and with a lot of benefits for both content creators and content users."

Greenberg, et al (2006) notes that most of the literature on the topic of automatic subject metadata generation falls under one of two categories. The first, "experimental research," is aimed primarily at the study of different techniques of information retrieval and the content of digital resources. The second is identified as "applications research," and concerns the development of tools, both in the form of applications for metadata generation and software designed for content creation. Greenberg identifies a

disconnection between the two as well as the need to consider both parts in conjunction with one another.

This revelation prompts one to question: How can information retrieval theory and techniques be used to create tools in order to satisfy the growing metadata needs of repositories, content creators, and users? In a 2006 survey of metadata experts, most were not comfortable with replacing human generation or evaluation of metadata with automatic applications; they did, however, agree for the most part that applications which automatically generate metadata should be used as tools which complement metadata generation performed by humans and enhance their expertise and time-effectiveness. Experts found it important that content standards be used along with algorithms in such applications (Greenberg, 2006). It has also been suggested, given the percentage of collections that are actually heavily used, that libraries and repositories identify the "most important" resources and allow humans to describe them while automatic generators describe the rest. For large databases and digital repositories, however, automatic indexing and metadata generation is widely seen as crucial to sustained success.

While some institutions or organizations may have the resources to rely solely on human generated metadata, the information explosion continues to make automated metadata generation more attractive and more crucial for both organization and retrieval of information. Indeed, even those with the luxury of adequate human indexers must see automated metadata generation as an attractive tool if nothing else; additionally, increase in metadata generation by content creators, users, and other non-professionals adds to the importance of developing quality automatic tools. Because of these realities, the discussion undertaken by Cardinaels and others as to how automatically generated

metadata compares in quality with human generated metadata is becoming increasingly irrelevant, and it is instead becoming more appropriate to ask: How can the quality of automatically generated metadata be optimized? What techniques and metadata generation methods will produce the highest quality metadata records with the least noise? This study seeks to compare two methods of automatic annotation and, in doing so, contribute to the body of information which may lead us closer to these answers.

RESEARCH OBJECTIVES

The research presented here addresses this need by exploring the quality of automatically generated metadata produced by two annotators using different tools and techniques. This study examines whether annotators using machine learning produce better quality results than those using basic term matching by testing the same set of documents on two automatic annotators. Quality, in this experiment, is examined across characteristics such as relevance, precision, specificity, and exhaustivity.

METHODOLOGY

In order to address the need for high quality metadata which is generated automatically, an experiment was conducted to compare the results produced by an annotator using basic term matching with those generated by an annotator using machine learning. For the purposes of this study, “annotators” will be defined as services which identify concepts in user-submitted texts, more specifically, the NCBO BioPortal application and the HIVE vocabulary server. Both annotators focus on the discipline of

evolutionary biology. "Machine learning" will be defined as the algorithms employed by HIVE vocabulary server which "teach" the system about concept relationships.

NCBO BioPortal

NCBO BioPortal is a web-based application into which ontologies can be uploaded and shared. The application was created by the National Center for Biomedical Ontology consortium (NCBO BioPortal, 2010). There are currently 194 ontologies in BioPortal, and these ontologies comprise a dictionary which includes well over one million terms. The ontologies are encoded in one of three formats: RRF, OBO, and OWL. Although semantic expansion occurs during the annotation process, the technique employed by BioPortal is basic term-matching.

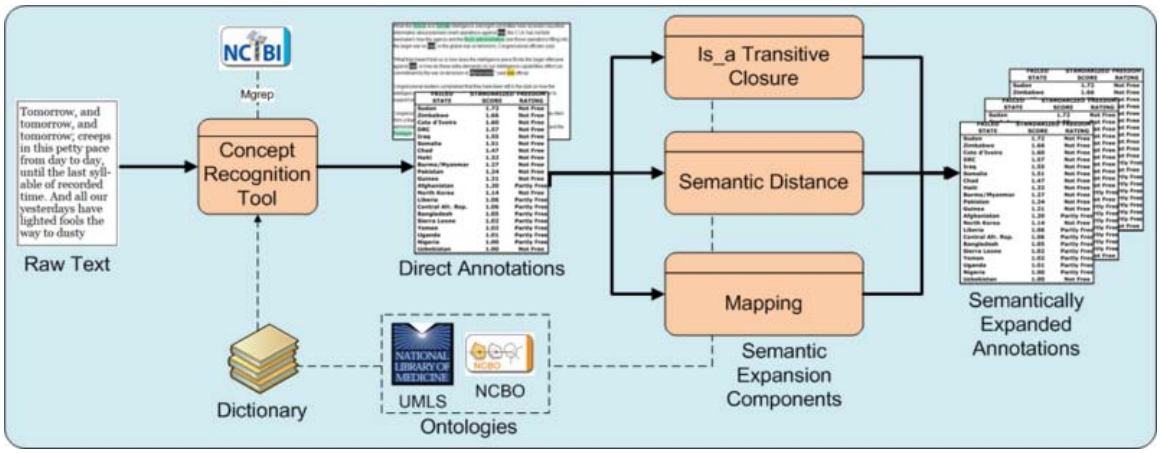


Figure 1. BioPortal Workflow

The chart above depicts the workflow of the NCBO BioPortal Annotator. There are three steps in the annotation process which produces the term sets:

1. The annotator extracts terms and phrases from the ontologies which are exact matches of those found in the text.

2. The annotator uses the is_a hierarchical relationship to expand the terms identified in the text and identify hierarchical parent-child relationships in the ontologies.
3. The terms selected are matched across the ontology mappings to extract terms from other ontologies which were identified manually at ontology ingestion.

HIVE (Helping Interdisciplinary Vocabulary Engineering) project

HIVE (Helping Interdisciplinary Vocabulary Engineering) is a collaborative effort between the Metadata Research Center in the School of Information and Library Science of the University of North Carolina at Chapel Hill and NESCent, the National Evolutionary Synthesis Center in Durham, NC (HIVE, 2008). HIVE is a project funded by the Institute of Museum and Library Services which uses controlled vocabularies along with Keyphrase Extraction Algorithm machine learning techniques in order to generate automatic subject metadata. The HIVE project was initiated to address challenges associated with the use of more than one vocabulary or thesauri in metadata generation, and is supported by research which indicates that multiple controlled vocabularies are necessary in order to accurately describe interdisciplinary subjects. The HIVE project's goal is to create an application which can bridge multiple discipline-specific controlled vocabularies and automatically generate subject metadata in order to aid information professionals, repository contributors, and other users in the description of interdisciplinary resources. The vocabularies used in HIVE are encoded in Simple Knowledge Organization Systems (SKOS).

The HIVE project is composed of three parts:

- *Building HIVE* seeks to provide a solution to the issues of affordability, efficiency, user friendliness and interoperability in the metadata creation process;
- *Sharing HIVE* involves continuing education for professional metadata creators in various settings to emphasize the value of emerging technologies in the use of multiple controlled vocabularies;
- *Evaluating HIVE* involves determining the effectiveness of HIVE in library, museum and archival environments as well as within Dryad, a digital repository which stores data objects related to published research.

In addition to the use of controlled vocabularies, HIVE employs algorithms which "teach" the system about concept relationships. HIVE uses KEA++ automatic metadata extraction techniques which include a training phase and a testing phase. Ramon Perez Aguera (2009) explains the KEA++ teachniques employed in HIVE; the training phase involves the use of controlled vocabularies by human indexers to index documents and create a training set. In the testing phase, the application attempts to use the training set to index documents of similar makeup.

In addition to the training set, other features include:

- *A term frequency/ inverse document frequency (TF/IDF) weight* which assesses keyphrase relevance by comparing the occurrence of a keyphrase in the document to the occurrence of that keyphrase in the entire document set.
- *A "first occurrence" weight* which assesses the importance of the keyphrase in a document according to the place in the document where the keyphrase first occurs.

- *A "node degree" feature* which considers the relationship of the keyphrase to the structure of the thesaurus; the degree is determined by the relationship of the keyphrase to thesaurus terms, other keyphrases, and a ratio of the two.

HIVE anticipates that the use of these machine learning features makes possible the generation of automatic subject metadata which has greater relevance, less "noise," and matches more closely the metadata creation that would result from manual indexing by humans. The problem-solving process involving the training set and other features is intended to help HIVE "learn" about concept relationships and thus improve subject metadata generation beyond the rigidity of simple term-matching schemes.

Document set

The document set used in this research was drawn from the partner list of Dryad data repository. The Dryad data repository stores data object that underlie published research in the field of evolutionary biology. For example, consider a research article; it may have items such as histograms, pie charts, or tables which summarize data used in the research. The Dryad project stores the data such as the files or other raw data which underlie the published result, but not the published article. Because creating metadata for each data object is time consuming, the published research article serves as a source, despite the fact that there may not be a one-to-one correspondence. In other words, the abstract for the published article containing the data is used to generate an automatic annotation for the data. In the case of published articles that also have keywords, they are used to create metadata as well.

Procedures

This project tested HIVE and NCBO BioPortal for a selection of Dryad article abstracts, and included the following steps:

1. Selection of document set

The Dryad data repository partner list includes twelve journals related to evolutionary biology. Using a research randomizer to ensure indiscriminate selection, ten of twelve journals were selected from the partner list. Two articles were randomly chosen from the most recent issue of each journal; the abstracts of the articles were then extracted for automatic subject metadata generation by the HIVE and NCBO BioPortal annotators.

2. Automatic subject metadata generation

After the document set was selected, the abstract for each selected article was uploaded into both the HIVE and NCBO BioPortal annotators for automatic subject metadata generation. The annotation results for each document were captured for evaluation.

3. Evaluation

The recorded results will then be used to evaluate the quality of results produced by each annotator for each document. For the purposes of this project, the results of best quality are those with higher numbers of unique relevant terms, greater precision, higher specificity of terms, and higher levels of exhaustivity per set. The quality of the results produced by each annotator was measured according to the following characteristics:

- *Specificity*- For the purposes of this project, specificity describes the extent to which terms produced describe the resource accurately and precisely. Specificity of terms was measured on an ordinal scale with values of 1 (good), 2 (fair), 3 (poor). Specificity values were determined according to the judgment of each evaluator.
- *Exhaustivity*- For the purposes of this project, exhaustivity is defined as the extent of coverage a term set presents. Exhaustivity of result sets was also measured on an ordinal scale with values of 1 (good), 2 (fair), 3 (poor). Exhaustivity values were determined according to the judgment of each evaluator.
- *Relevance*- For each set of results, human evaluators identified unique relevant terms produced by the annotator. Terms were defined as relevant by each evaluator according to their personal assessment.
- *Precision*- After relevant terms for each term set produced by each annotator were identified, counted, and recorded, a basic precision measure was calculated by dividing the number of relevant terms by the total number of terms.

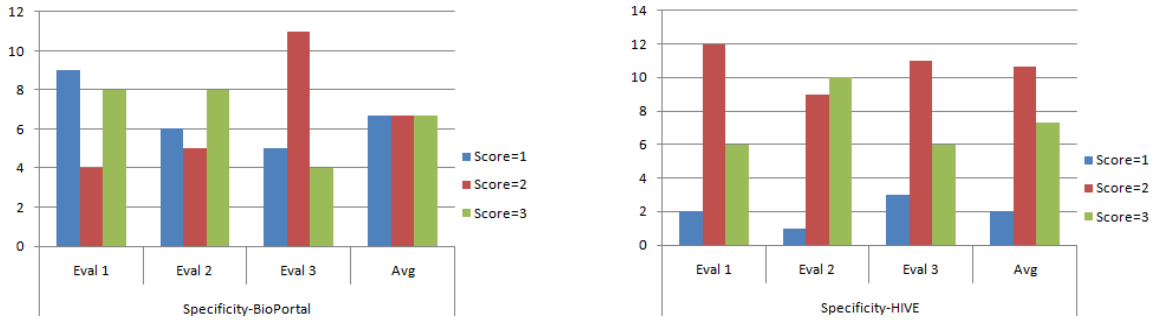
Relevance, specificity, and exhaustivity evaluations for this study relied on human assessment. To avoid single evaluator biases, all of the metadata results from both systems were evaluated by three people: the researcher, and two independent evaluators- all of whom are knowledgeable in the area of metadata and knowledge organization. After the evaluators assessed the term sets produced by each annotator, the averages of these results were taken, minimizing any potential bias which might result from the subjective nature of the characteristics measured.

RESULTS

Following the selection of the document set, annotations of the selected abstracts were automatically generated by both the HIVE and NCBO BioPortal applications. The annotations were then appraised by three human evaluators, the results of which were recorded, averaged, and analyzed. This section presents key of results regarding the mean specificity, exhaustivity, relevant terms, and precision for each annotator. Additionally, this section includes a discussion which compares the individual results produced by the three human evaluators and identifies possibilities with regard to the impact of human subjectivity on this study.

Specificity

For each annotator, the mean of the scores for the document set reported by each evaluator was calculated; the mean for each of the three evaluators was then averaged to produce an overall specificity rating.



Figures 2 &3. Specificity (by evaluator)

In terms of average specificity, the HIVE annotator outscored the NCBO BioPortal annotator by a score of 2.3 to 2, respectively. Once averaged, the percentage of

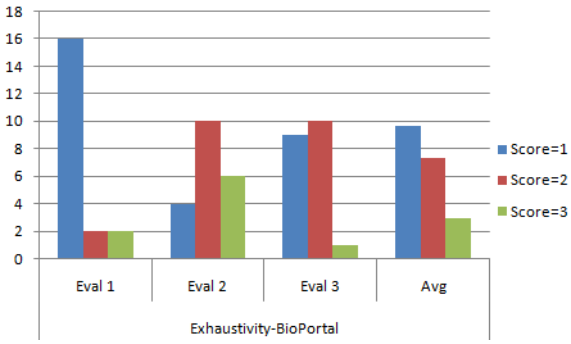
documents receiving each score (1-3) was calculated. Average scores for result sets produced by the NCBO BioPortal annotator were split evenly across the spectrum at 33.33%, meaning that the NCBO BioPortal annotator produced as many term sets with "poor" specificity as it did "good" specificity. This indicates that the quality of specificity found in NCBO BioPortal result sets is rather inconsistent.

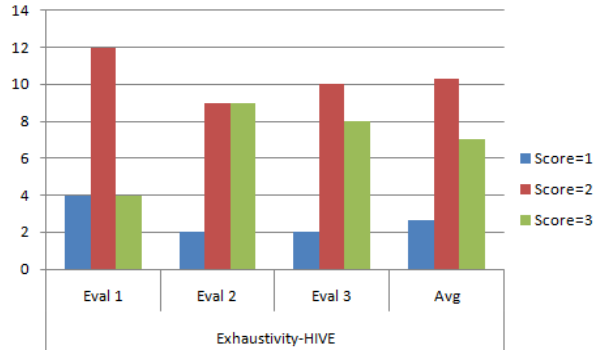
The scores for the HIVE result sets were considerably more skewed. A mere 10% of results sets were given a score of 1, while the majority of results sets (53.33%) were assigned a score of 2, and 36.67% of result sets were assigned a score of 3. This indicates a slightly higher probability that HIVE will produce results which are considered "good," and a more significant (23.34%) probability for result sets to be considered "fair" or better than those produced by NCBO BioPortal.

Exhaustivity

Like specificity, the mean exhaustivity for each score set as assigned by each evaluator was calculated; the mean exhaustivity according to each evaluator was then averaged to determine overall exhaustivity. In terms of exhaustivity, NCBO BioPortal was outscored by HIVE once again; the overall exhaustivity of NCBO BioPortal results was 1.7, while HIVE result sets averaged a score of 2.2. The breakdown of exhaustivity scores assigned to NCBO BioPortal term sets was much more disparate than that of the specificity scores; the majority of the term sets produced by NCBO BioPortal (48.33%) were assigned a score of 1, while 36.67% received a score of 2, and 15% received a score of 3. While these percentages of exhaustivity scores for NCBO BioPortal term sets are certainly more consistent than the specificity scores, they unfortunately indicate that the

exhausitivity of NCBO BioPortal annotations are most likely to be poor, and much more likely to be fair than good.





Figures 4 & 5. Exhaustivity (by evaluator)

The breakdown of HIVE score percentages with regard to specificity was quite similar to the exhaustivity scores. A specificity score of 2 was assigned to HIVE results most often (51.67%), with 35% percent receiving a score of 3, and only 13% receiving a score of 1. This is remarkable when compared with NCBO BioPortal results; the 51.67% of HIVE results which received a score of 2 is equal to the sum of the percentages of NCBO BioPortal results which received a score of 2 *or* 3. The HIVE annotation sets were 35% more likely to receive a "fair" or "good" rating than producing higher quality result sets in terms of exhaustivity.

Relevant terms and Precision

The average number of relevant terms produced by each annotator varied. The average number of relevant terms per set for NCBO BioPortal annotations ranged from 2.7 to 8.3, while the average number of relevant terms per set for HIVE ranged from 3.3 to 10.6. Overall, NCBO BioPortal produced an average of 5.6 relevant terms per

document, and the HIVE annotator produced a slightly higher average of 5.78 terms per document. This average becomes more significant when considering the fact that NCBO BioPortal produces an incredibly higher number of total terms per document than HIVE, an approximate ratio of 1:16.

Precision scores for each annotator were calculated by dividing the average number of relevant terms produced by the total number of terms produced for each document by each annotator. The mean of these averages was calculated in order to determine an overall precision rate for each annotator. The overall precision rate for NCBO BioPortal term sets was .015, and the HIVE term sets had an overall precision of .209. That the overall precision of the HIVE annotator is more than sixteen times that of the NCBO BioPortal annotator is quite remarkable considering the small number of thesauri HIVE uses to generate the annotations in comparison to the vast number of ontologies employed by NCBO BioPortal.

Evaluator Comparisons and Impact

The number of relevant terms identified by the three evaluators varied, as did the scores assigned for specificity and exhaustivity. For example, Evaluator 1 assigned far more NCBO BioPortal term sets score of 1 than either the other two evaluators for both specificity and exhaustivity. Evaluator 3 assigned more scores of 2 in both categories (specificity and exhaustivity) for both annotators. This is to be expected, however, as the assessment of these characteristics is quite subjective. It is worth noting, however, that despite variance among the evaluators on individual term sets, that individual averages of all three evaluators rated HIVE higher than NCBO BioPortal in terms of both

exhaustivity and specificity. This lends some weight to indication of HIVE's superiority in these areas, because it indicates that despite the inconsistency of indexer discretion, the same conclusion was reached by all three evaluators, albeit in slightly different degrees.

One individual average, Evaluator 1's relevant terms per set, was inconsistent with the overall average; Evaluator 1 identified more relevant terms per set in the HIVE annotations, with a 4.8 average for HIVE results and a 3.4 average for NCBO BioPortal. Additionally, the average relevant terms per set identified by Evaluator 1 was significantly lower for both annotators than either of the other two evaluators. Because these averages are used to create the overall mean which is used in the precision measurement, a difference in the individual average of one evaluator could cause significant change in precision. However, if the averages of Evaluator 1are discarded, the precision measurements for each annotator become 0.015 for NCBO BioPortal and 0.227 for HIVE, which still would mean that HIVE term sets are on average fifteen times more precise than that of NCBO BioPortal.

DISCUSSION

The evidence overwhelmingly suggests that term sets produced by the HIVE annotator are superior to those produced by the NCBO BioPortal application. Though two evaluators found that on average the NCBO BioPortal annotator produces more relevant terms per set than HIVE, the fact that the total term ratio is 16:1 means that HIVE's term sets are drastically lower in noise. Additionally, when considering these results with attention to the low number of vocabularies used by the HIVE annotator in comparison with the NCBO BioPortal, there seems be an indication that the machine

learning algorithms which are the power behind HIVE's annotator outperform the basic term-matching scheme of NCBO BioPortal.

The sheer number of ontologies employed by the NCBO BioPortal annotator would seem to suggest that its results would have a greater probability of being more exhaustive than that of the HIVE annotator, which uses only 3 vocabularies; surprisingly, this was not the case. This could be a result of the structural rigidity of the ontologies used in NCBO BioPortal. Another possible explanation for these results could be that while NCBO BioPortal uses a higher number of ontologies, the ontologies used could have more similar or shared representations of concepts than that of the HIVE vocabularies; if this is the case, then these representations shared by multiple ontologies could simply lead to higher repetition of the same terms rather than greater exhaustivity. Such repetitions would explain to some extent the drastically lower precision rating of the NCBO BioPortal in comparison to HIVE, because relevance was measured by unique terms only.

Overall, the results of this experiment which indicate that HIVE is a superior tool in terms of precision, specificity, and exhaustivity are not unexpected. Because it employs machine learning techniques, the HIVE annotator is able to distinguish concepts to some extent that are not actually present as terms in the document. Additionally, HIVE uses algorithms which can distinguish the importance of the concepts found within the document, which is likely to increase the relevance of terms produced by HIVE.

LIMITATIONS

There are several limitations to this study. Perhaps the most apparent limitation of this experiment is that it tested only two automatic annotators of the many which exist. It is likely that results produced by different annotators, both those using basic term matching and machine learning, will differ from those produced by NCBO BioPortal and HIVE. Different combinations of machine learning algorithms or mapping schemes between vocabularies and ontologies could lead to altered results. Additionally, the use of different ontologies and vocabularies themselves in the annotation applications would likely result in significant differences.

Other limitations involve the document set used in the study. For instance, it is likely that certain document types may be more suited than others for a particular type of annotation. In this vein, document sets from disciplines other than evolutionary biology may produce different results under these same annotation techniques. Resources in languages other than English, too, would present new issues to be explored for machine learning annotators in particular, as the hierarchical structures of language vary.

Because HIVE uses controlled vocabularies, while NCBO BioPortal uses ontologies whose structures are more formal and rigid, this adds another variable to the experiment which may affect the results of both annotators in addition to the machine learning or term-matching schemes. Future comparisons in basic term-matching and machine learning annotations would be better served to use two annotators which employ the same vocabularies or ontologies to produce their annotation.

Though three evaluators were used to strengthen validity, the identification of relevance, specificity, and exhaustiveness are highly subjective. Many studies assessing

relevance have used a “gold standard” created by a human indexer or metadata professional to assess these types of characteristics; however, even “gold standard” sets of manual annotations are limited by the fact that human indexers produce high quality but inconsistent annotations, and therefore while a repetition of this study with a “gold standard” set of annotations may produce different results, those results would still vary according to the humans used to create the annotations.

CONCLUSION

The study presented in this paper examined the quality of results produced by two different annotators, HIVE and NCBO BioPortal, in an effort to examine whether machine learning or basic term matching produces higher quality automatically generated metadata. The metadata generated by each annotator was evaluated in terms of relevance, precision, specificity, and exhaustivity. The term sets produced by the HIVE annotator outperformed the NCBO BioPortal annotator were on average 10 percent higher in terms of specificity, 17 percent higher with regard to exhaustivity, and had 19.4 percent higher precision scores than the results produced by BioPortal The results of this study indicate that machine learning may be a better technique for automatic metadata generation in the form of abstract annotation in terms of quality. Given the limited size and domain of the HIVE project’s current annotator, issues of scalability and interoperability between domains still present challenges for future research.

This area of research is quite significant, however, in that further developments in machine learning theory and its incorporation with automatic metadata generation applications will have meaningful impact on description, organization, and retrieval of

information. This potential impact could benefit various types of repositories and other organizations or individuals generating metadata. Users of information in many fields will also benefit from developments which enhance repositories' and other institutions' capability to describe and provide access to greater numbers of resources with higher quality and efficiency. Further, improvements in these areas will enhance interoperability between disciplines, which could in turn facilitate new research and the creation of new information with the potential to take knowledge in every discipline to new heights.

References

Aguera, J. (2009) How is working KEA, the HIVE automatic metadata extraction algorithm.

Anderson, J., & Perez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management, 37*(2), 231. Retrieved from Business Source Premier database.

Anderson, J., & Perez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing & Management, 37*(2), 255. Retrieved from Business Source Premier database.

Cardinaels, K., Meire, M., & Duval, E. (2005). Automating metadata generation: the simple indexing interface. In *Proceedings of WWW 2005: 14th international conference on World Wide Web,* ACM Press, 548-556.

Greenberg, J. (2004, in press). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging, 6(4): 59-82.

Greenberg, J., Spurgin, K. & Crystal, A. (2006). Functionalities for automatic metadata

generation applications: a survey of metadata experts' opinions. *International Journal of*

Metadata, Semantics & Ontologies, 1(1).

Helping Interdisciplinary Vocabulary Engineering. (2008). *HIVE Main Page.* Retreived from https://www.nescent.org/sites/hive/Main_Page.

Liddy, E. D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N. E., Diekema, A., McCracken, N. J., Silverstein, J., & Sutton, S. A. (2002). Automatic metadata generation & evaluation. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 11-15, 2002, Tampere, Finland New York: ACM Press, 401-402.

Marchionini, G., & Maurer, H. (1995). The Roles of Digital Libraries in Teaching and Learning. Communications of the ACM, 38(4), 67-75. Retrieved from Academic Search Premier database.

Miksa, F L. (1998). *The DDC, the universe of knowledge, and the post-modern library.* Albany, N.Y.: Forest Press.

NCBO BioPortal. (2010). *Annotator User Guide*. Retrieved from http://www.bioontology.org/wiki/index.php/Annotator_User_Guide.

O'Conner, J. Mechanized indexing methods and their testing. *Journal of the ACM,* 11(4):437-449, October 1964.

Salton, G. (1970). Automatic text analysis. *Science,* 168 (3929):335.

Shen, D., Yang, Q., & Chen, Z. (2007). Noise reduction through summarization for web-page classification. *Information Processing and Management, 43*(6), 1735-1747. doi:10.1016/j.ipm.2007.01.013