Computational Curation and the Application of Large-Scale Vocabularies

Sam Grabus
MRC, College of Computing & Informatics
Drexel University
Philadelphia, USA
0000-0003-4670-5690

Jane Greenberg
MRC, College of Computing & Informatics
Drexel University
Philadelphia, USA
0000-0001-7819-5360

***Abstract*—Paper presents an exploratory case study comparing stemming and lemmatization results for the automatic application of large-scale controlled vocabularies processed against archival encyclopedia entries. The results report relative recall and precision evaluations across both results. Research shows that while stemming has a higher relative recall, lemmatization results in a higher relevance score and eliminates the over-stemming challenges. Results provide insight into improving automatic curation workflows for archival resources.**

***Index Terms*—controlled vocabularies, stemming, lemmatization, natural language processing (NLP), automatic curation**

I. Introduction

This digital transformation of historical records has inspired multiple innovative disciplinary trends, including digital humanities and the focused area of computational archival science [1] [2]. Archives, as historical records, are resources generated through daily activity. Records are preserved as archives, given their *enduring and evidential value* [3]. Digital preservation, further, allows the archives to be transformed into computational-ready formats. As a result, researchers can mine these archives and apply empirical analysis on a scale that far exceeds what was possible with the analog artifact. Such approaches can lead novel findings and support a new understanding of our past (e.g., [4]). Common approaches for mining digital archival content include natural language processing (NLP), named entity recognition (NER), clustering, topic maps, and more recently exploration of knowledge graphs (e.g., [5]) and neural networks (e.g., [6]). These approaches are not necessarily exclusive, and in some cases, build on one another. Another area of importance that has generally garnered less attention is the automatic application of structured vocabularies—specifically, controlled vocabularies. This is somewhat surprising, given that the development of linked open data (LOD) terminologies has continued to advance, with major reach institutions, such as the U.S. Library of Congress making all of their controlled vocabularies accessible as LOD. One reason for the limited application may be traditional perceptions about the limitations of controlled vocabulary systems in the manual environment [7]. Another likely reason is limited knowledge about the best way to apply these systems. The is particularly striking when working with historical text, where the *literary warrant*, language of the archival text does not align with end-user warrant, the language of the current day researcher/archival system user [8]. This predicament underscores the need to explore if LOD controlled vocabularies, ontologies, and similar semantic systems are useful for exploring digital archives; and, more precisely, what approaches seem to be the most effective.

These questions have motivated researchers affiliated with both the 19th-Century Knowledge Project and Drexel University's Metadata Research Center to explore the use of LOD controlled vocabularies for automatically indexing archival resources. A specific aim is to gain an understanding of what approaches may work best for addressing linguistical challenges, such as *anachronistic terminology* or *dialectal variants* (e.g., American and British English). This research presented in this paper, seeks to address these questions and reports on work exploring the automatic application of the *Library of Congress Subject Headings* (LCSH) for assigning subject (topical) metadata to archival encyclopedia records. The reach compares stemming and lemmatization applying both the 1910 LCSH and the FAST-LCSH (2020), to archival entries in the 7th edition of the *Encyclopedia Britannica*, published 1842. The sections that follow provide background information on text normalization—specifically stemming and lemmatization; automatic indexing with controlled vocabularies and ontologies; followed by our research aim, methods, results, and a discussion. The conclusion highlights our key results and notes next steps.

II. Text Normalization

Text normalization is a pre-processing step used to transform terms into a standardized form prior to computational use. Normalization typically includes removing punctuation, contractions, and stop-words, transforming the text into lowercase, tokenization, and applying either stemming or lemmatization. Both stemming and lemmatization are used to improve information retrieval results, but there are important differences that can inform the decision to use one over the other.

Stemming typically transforms terms to their root form by removing the derivational affixes from the end of the word [9]. For example, the Porter Stemmer transforms the term *Mountaineering* to *mountain*. By transforming terms with the same root to the same lexical unit, stemming increases information

retrieval recall. This approach is particularly advantageous when compensating for sparse data [9]. Lemmatization is a dictionary-based approach to term standardization that preserves word sense of a term by transforming it to the dictionary form and only removing inflectional endings [9]. For example, NLTK's WordNet Lemmatizer transforms *Mountaineering* to *mountaineer*. Terms with the same root would be transformed to different lexical units to preserve the word sense, which can improve information retrieval quality.

III. Automatic Indexing with Controlled Vocabularies/Ontologies

Several information retrieval researchers have studied the use of lemmatization of the input text and/or vocabulary terms when using controlled vocabularies or ontologies to annotate documents. Sinkkilä et al. [10] compared three stemming and lemmatization algorithms and found that the two lemmatization algorithms performed significantly better than the stemmer, with indexing quality comparable to human indexing. When comparing Bulgarian language results with stemming and indexing, Borisova [11] found that stemming did not allow for "consistent interpretation of the common words". One key consideration when indexing with a vocabulary is that the stemming process can lead to what Sinkkilä et al. [10] describe as "imperfect morphological analysis," with a stemmed document word connecting to a different word that has the same stemmed form. This has been referred to as over-stemming, which can decrease the precision of information retrieval results and "degrade classification performance" [9]. For example, for the six FAST-Topical controlled vocabulary terms in Table 1 below, the Porter Stemmer returns the same lexical unit, which removes any indication of word sense differences. Regardless of semantic and contextual word sense, each of these terms would be ranked equally during the indexing process because their normalized form is identical.

TABLE I
Examples of text normalization for FAST-Topical controlled Vocabulary terms using the Porter stemmer and the NLTK WordNet lemmatizer.

FAST-Topical Vocabulary Terms	Porter Stemmer	NLTK WordNet Lemmatizer
Capital	Capit	Capital
Capitalism	Capit	Capitalism
Capitalization	Capit	Capitalization
Capitellida	Capit	Capitellida
Capitols	Capit	Capitol
Capitulations	Capit	Capitulations

Some researchers have normalized both the input text and the vocabulary terms to which they will be matched. Martinez-Romero et al. [12] compared two versions of the NCBO Ontology Recommender, with the second version lemmatizing both the input terms and the dictionary terms. Their results found that the lemmatized approach provided higher-quality suggestions, better coverage, more detailed information about the concepts, increased specialization for the input data, and greater acceptance and use of the ontology in the biomedical community [12].

IV. Research Aims

This overall research aim of this work is to explore if LOD controlled vocabularies are useful gaining insight into the content of digital archives. More precisely, we seek to explore computational approaches seem to be the most effective. The key objective is to compare stemming and lemmatization.

V. Methods and Procedure

To address our aims, we pursued an exploratory case study method, with a sample of archival digitized entries was drawn from 7th edition of the *Encyclopedia Britannica*, and they automatically indexing using the *1910 LCSH* and the *FAST-Topical LCSH* vocabularies. Both are large-scale general domain vocabularies. The 1910 LCSH has 26,780 concepts, and the FAST-Topical has 460,110 concepts. The encyclopedia entries were processed through the Helping Interdisciplinary Vocabulary Engineering (HIVE) technology, which supports automatic metadata generation using multiple controlled vocabularies [13] Figure 1). The protocol for comparing the indexing output using each normalization approach will focus on two key steps:

a) Relative Recall: Using a convenience sample of 3,182 encyclopedia entries, automatic indexing results were generated using both stemming and lemmatization. Two controlled vocabularies were used in combination: *1910 LCSH* and *2021 FAST-Topical*. Relative recall was assessed across the output.

b) Precision Evaluation: A random sample of 50 entries was selected from the larger convenience sample. Using a four-tierd relevance scale from 0 (not relevant) to 3 (highly relevant), a human evaluator ranked the relevance of each term across the indexing output as it related to the context of the encyclopedia entry. The portion of relevant terms were determined, along with average difference in discounted cumulative gain, over-stemming occurrences, and homonym occurrences.

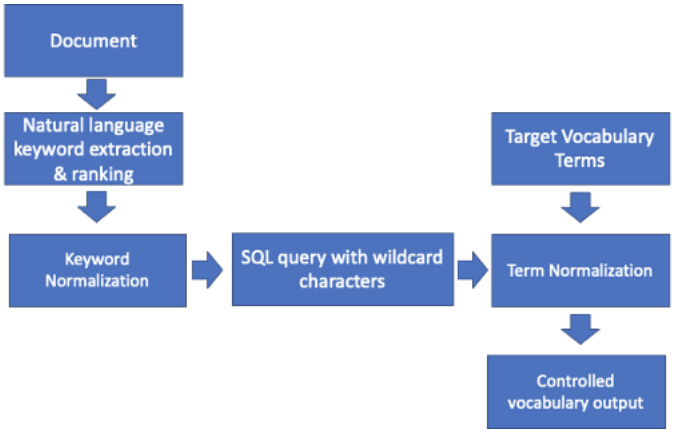

Fig. 1. Overview of the HIVE workflow

VI. Results

Our analysis covers the stemming and lemmatization results for the subject headings generated using HIVE. We report specifically on two aspects: 1) The relative recall; and 2) The relevance of the indexing output terms in relation to the indexed document, as determined by a human evaluator. Table 2 provides the number of terms retrieved and relative recall for the two normalization approaches. The relative recall when using the stemming approach is 52 percent, and 48 percent using lemmatization.

TABLE II
Relative recall for Stemming and Lemmatization

Search Query	Stemming		Lemmatization	
	Number of Terms	*Relative Recall*	*Number of Terms*	*Relative Recall*
Entry 1	12	0.52	11	0.48
Entry 2	11	0.5	11	0.5
Entry 3	2	0.5	2	0.5
Entry 4	5	0.71	2	0.29
Repeated for all 3182 IR queries	...	...	...	...
Total	34,135	**0.52**	31,337	**0.48**

[a]Total across both approaches: 65,472.

The smaller random sample of fifty entries was then evaluated for semantic relevance to the context of the indexed document. Table 3 below provides a high-level view of the indexing output and relevance rankings. Relevant terms are presented two ways: 1) conservatively, including only terms that were ranked three or two, and 2) more inclusively, including terms that were ranked three, two, or one.

Relevant results accounted for between 65 and 78 percent of the indexing output when using the stemming approach. Using the lemmatization approach, relevant results accounted for between 75 and 90 percent of the indexing output. Discounted Cumulative Gain by entry increased by an average of .32 using lemmatization over the stemming approach. Homonym indexing errors occurred equally using each normalization approach, accounting for approximately 7 percent of the indexing results. Over-stemming occurs in 11 percent of the total stemming output, while the issue does not occur using lemmatization.

TABLE III
Relevance evaluation comparing indexing output using stemming vs. lemmatization

		Normalization Approach	
		Stemming	***Lemmatization***
Total Terms Retrieved		483	444
1910 LCSH Terms		235	220
2020 FAST-Topical Terms		248	224
Relevance Rankings			
Relevant	3 and 2	316 (65.42%)	334 (75.22%)
	3,2,1	379 (78.47%)	398 (89.64%)
Non-Relevant	1 and 0	167 (34.58%)	110 (24.78%)
	Just 0	104 (21.53%)	46 (10.36%)
Average DCG difference by Entry		-0.32	+0.32
Over-stemming Occurrences		53 (10.97%)	N/A
Homonym Occurrences		34 (7.04%)	31 (6.98%)

Table 4 below provides examples of over-stemming from the results. For instance, when indexing an encyclopedia entry about Tibet, stemming caused the word *Mountains* to be matched to *Mountains*, *Mountaineers*, and *Mountaineering*, with *Mountains* being the only relevant match.

TABLE IV
Examples of indexing output differences for specific words using stemming and lemmatization. Bolded terms indicate relevant output, as determined through examining the context in which the term was used in the encyclopedia entry

Term in Indexed Document	Context	Indexing Output: Stemming	Indexing Output: Lemmatization
Communes	SAINTES , an arrondissement of the department of the Lower Charente, in France. It is 595 square miles in extent, and comprehends eight cantons, divided into 109 communes, with 184,871 inhabitants in 1836. The	Communities Communication Communism **Communalism**	N/A
Capital	RAAB , a city of the Austrian kingdom of Hungary, in the province of the hither Danube, the capital of a circle of the same name, and of the mouth of the Raab. It is well built,	**Capital** Capitalization Capitalism	**Capital**
African	his life. Rabirius escaped with difficulty from Egypt; and at his return to Rome he was accused by the senate of having lent money to an African prince for unlawful purposes. He was ably	**Africans** Africanization	**Africans**
Mountains	rise. Tibet is separated, about the twenty-eighth degree of north latitude, from Bootan by the Sumoonang Mountains, part of the great Himalaya chain ;	**Mountains** Mountaineers Mountaineering	**Mountains**

VII. Discussion and Conclusion

The research above presents a comparison of automatic indexing results when using stemming or lemmatization to normalize the input text and controlled vocabulary terms. Lemmatization demonstrated a 10-11.17 percent increase in precision over the stemming approach, an average .32 increase in discounted cumulative gain for each entry and eliminated the over-stemming challenge. While the stemming approach demonstrates greater overall recall, the lemmatization resulted in higher precision, which is ultimately better for subject representation of historical archives. This result may assist the development of automatic curation workflows, where curators and other staff seek to apply controlled terminologies to assist with resource discovery and collection collocation. Most importantly, the results provide insight into the best approach for the automatic application of large-scale vocabularies to generate high-quality metadata, and help address the curation bottleneck. Today, the majority of digital archives still lack consistent subject descriptors, and this impacts discovery and use. The results are consistent with other studies that compare stemming and lemmatization (e.g., [10]. Overall, the results offer empirical evidence for archival collections, particularly in collections that are general domain in which stemming results in inaccuracies. Finally, the results have informed next steps with making historical encyclopedia entries accessible, where

we are applying lemmatization, working with both the *1910 LCSH* and *FAST-Topical*.

ACKNOWLEDGMENT

We'd like to acknowledge Professor Peter Logan and Sr. Software Developer Joan Boone for the infrastructure contributions that made this research possible.

REFERENCES

[1] R. Marciano, S. Agarrat, H. Frisch, M. R. Hunt, K. Jain, G. Kocienda, H. Krauss, C. Liu, M. McKinley, D. Mir, C. Mullane, E. Patterson, D. Pradhan, J. Santos, B. Schams, H. S. Shiue, A. J. Silva, M. Suri, T. Turabi, M. Vasselli, and J. Xu, "Reframing Digital Curation practices through a computational thinking framework," 2019 IEEE International Conference on Big Data (Big Data), 2019.

[2] N. Payne, "Stirring the cauldron: Redefining computational archival science (CAS) for the Big Data Domain," 2018 IEEE International Conference on Big Data (Big Data), 2018.

[3] T. R. Schellenberg, Modern Archives: Principles and techniques, by T.R. Schellenberg. Chicago, IL: University of Chicago Press, 1975.

[4] T. Underwood, Distant Horizons: Digital Evidence and Literary change. Chicago and London: The University of Chicago Press, 2019.

[5] X. Wang, R. Wang, Z. Bao, J. Liang, and W. Lu, "Effective Medical Archives processing using knowledge graphs," Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.

[6] T. Blanke, M. Bryant, and M. Hedges, "Understanding memories of the Holocaust—a new approach to neural networks in the Digital Humanities," Digital Scholarship in the Humanities, vol. 35, no. 1, pp. 17–33, 2019.

[7] T. Gross, A. G. Taylor, and D. N. Joudrey, "Still a lot to lose: The role of controlled vocabulary in keyword searching," Cataloging & Classification Quarterly, vol. 53, no. 1, pp. 1–39, 2014.

[8] F. W. Lancaster, Vocabulary Control for Information Retrieval. Arlington Va.: Information Resources, 1992.

[9] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval. Cambridge: Cambridge University Press, 2018.

[10] R. Sinkkilä, O. Suominen, and E. Hyvönen, "Automatic semantic subject indexing of web documents in highly inflected languages," Lecture Notes in Computer Science, pp. 215–229, 2011.

[11] N. Borisova, "An approach for ontology based information extraction," Information Technologies and Control, vol. 12, no. 1, pp. 15–20, 2014.

[12] M. Martínez-Romero, C. Jonquet, M. J. O'Connor, J. Graybeal, A. Pazos, and M. A. Musen, "NCBO ontology recommender 2.0: An enhanced approach for biomedical ontology recommendation," Journal of Biomedical Semantics, vol. 8, no. 1, 2017.

[13] J. Greenberg, X. Zhao, M. Monselise, S. Grabus, and J. Boone, "Knowledge Organization Systems: A network for AI with helping Interdisciplinary Vocabulary Engineering," Cataloging Classification Quarterly, pp. 1–20, 2021.