Automated Topic Analysis for Restricted Scope Health Corpora: Methodology and Comparison with Human Performance

Anthony Maeder
Flinders University, Australia
anthony.maeder@flinders.edu.au

Jennifer Tieman
Flinders University, Australia
jennifer.tieman@flinders.edu.au

Berth Naveda
Flinders University, Australia
bertha.naveda@flinders.edu.au

Stephanie Champion
Flinders University, Australia
stephanie.champion@flinders.edu.au

Tamara Agnew
Flinders University, Australia
tamara.agnew@flinders.edu.au

Abstract

This paper addresses the problem of identifying topics which describe information content, in restricted size sets of scientific papers extracted from publication databases. Conventional computational approaches, based on natural language processing using unsupervised classification algorithms, typically require large numbers of papers to achieve adequate training. The approach presented here uses a simpler word-frequency-based approach coupled with context modeling. An example is provided of its application to corpora resulting from a curated literature search site for COVID-19 research publications. The results are compared with a conventional human-based approach, indicating partial overlap in the topics identified. The findings suggest that computational approaches may provide an alternative to human expert topic analysis, provided adequate contextual models are available.

1. Introduction

Extracting useful information from scientific publications is a widespread need in research. It is usually accomplished by conducting systematized searching of peer-reviewed publication database sources using search expressions composed of distinct concepts describing the domain of interest, followed by an appropriate formal literature review process applied to the resulting papers [1]. Typically this requires involvement of human experts, with considerable time and effort expenditure, and is sensitive to their expertise. In certain instances, such as scoping reviews or trend analysis, the human experts are tasked with synthesizing a set of commonly occurring distinctive and prominent topics within the corpus. Automation of this undertaking using computational tools would offer gains in efficiency and repeatability, and consequently has attracted much recent research attention e.g. [2].

Text mining research has developed various methods for topic analysis and topic modeling, where a 'topic' is defined as "a subject within a text, represented by means of a cluster of words that are closely related to a seed word" [3]. Prominent topics detected by text mining may differ substantially from the original set of concepts which were used to define the scope of the corpus, in the construction of literature search expressions or for defining inclusion/exclusion criteria. These topics would generally provide further detail of the secondary focus areas covered in the corpus the papers, beyond the primary search focus. Commonly this task is achieved by semantic analysis approaches, involving iterative or probabilistic modeling techniques [4]. These approaches depend on large datasets of text which enable robust models to be constructed, whereas methods specifically for corpora of limited size and scope are still in their infancy [5].

An alternative approach is 'keyword analysis' which involves extracting and selecting highly relevant words or phrases from the title, abstract, or main body of text in the paper, based on their 'keyness' as determined by choice of "metrics of effect size and statistical significance" [6]. This is a process for which computational methods have been considered feasible but lacking a clear optimal method [7]. Reported results for different keyness models, different methods of keyword identification, and different underlying corpora, vary widely due to influence of semantic structure [8]. However, simple statistical analysis using

URI: https://hdl.handle.net/10125/70706
978-0-9981331-4-0

term frequency metrics has been shown to produce reasonable performance on research papers [9], comparable with human generated results [10].

There are several direct uses for topic analysis, including summarisation of coverage, defining major clusters of related work, assessing current trends, and tracking changes in focus over time. In many areas of health services research where there is inherently considerable diversity due to the highly inter-professional and multi-process environment (e.g. multiple care provision components; complex service delivery settings; compound clinical conditions), topic analysis could be useful for identifying different subsets of a given body of literature, which may be of specific relevance to different audiences. It may also be useful for tracking new or evolving areas of current interest in externally influenced circumstances (e.g. strategic planning inputs; assessing responses to policy reforms). These aspects are all necessities for maintaining up-to-date content for display on health information portals, or for making decisions on changes in emphasis and inclusion of new areas of material in health information repositories.

It is therefore of much interest whether topic analysis can be streamlined through partial automation, without loss of integrity of the 'gold standard' approach using human expert judgement and human-based knowledge synthesis. The underlying purpose of the work described here was to identify sets of major topics to describe content in sets of publications which had been extracted by professionally constructed search formulas derived by specialist librarians and knowledge management experts [11]. This work forms part of the ongoing 'Flinders Filters' program for curation of standardized search formulas to provide a clinical reference facility: further details are available at https://www.flinders.edu.au/flinders-digital-health-research-centre/flinders-filters .

The primary objective of this research was to identify topics in the Ageing and Aged Care knowledge domain generated from a set of recently constructed publicly accessible search filters, which provide COVID-19 related research papers containing clinical evidence of applicability and efficacy, from published scientific literature. The findings will be used to inform evidence retrieval for inclusion in web content for the End of Life Directions for Aged Care (ELDAC) project, designed to support palliative care and advance care planning in aged care as described at https://www.eldac.com.au/ . The search filters used to create the datasets for this work are available at http://oneclicksearching.com.au/ViruSearch/search and return newly-searched PubMed results in real-time. The secondary objective was to compare two approaches to addressing this problem, using respectively a human consensus-based approach and computational approach based on Natural Language Processing (NLP). This would contribute evidence for evaluating the feasibility of developing computational tools to automate parts of the topic generation process.

2. Methodology

Papers for both the human-based and computation-base methodological arms were sourced from three specific defined corpora selections related to Ageing and Aged care which were available from the above COVID-19 Evidence Link website: "Residential Aged Care", "Older People (>=65 years)", and "Isolation". These search filters were chosen as being sufficiently distinct from one another to provide opportunity for different topics to be identified for each corpus, while also returning a sufficiently small number of papers (between 50 and 100 in each case) to be feasible for application of both human-based and computational analysis. A fourth corpus with anticipated broader coverage of topics was sourced from the "General information" search option, to allow contextual comparison with the more specific corpora. All these searches were undertaken with the underlying constraints of "COVID-19" and "English language only" selected as settings.

The sets of papers for each corpus were sourced from the above website using the categories as described above, at a fixed timepoint (30 April 2020). A second set of papers in the first category (Residential Aged Care) was extracted at a later timepoint (5 June 2020) to enable change analysis to be performed and duplicate papers were excluded from the second corpus in that case. As the term "COVID-19" was incorporated as an AND clause in all search expressions, only papers published since Dec 2019 were noted to be included, as expected. The search filters produced compound PubMed search expressions which were used to extract a dataset of Full Text papers for each corpus. Due to imprecision in the PubMed internal data, some hits returned only the Titles of paper, some contained only Abstracts and no Full Text papers, and some entries were duplicated. All such cases were removed from the corpora that were used for further work in the project. As these constituted less than 10% of the total number of papers returned in the Pubmed searches, their omission was deemed to be negligible in effect.

2.1. Human-based method

An online human-based recommendation process for 'keywords' was established using the Covidence software package at https://www.covidence.org which is designed for conducting systematic literature reviews. Each corpus of Full Text papers was reviewed independently by three researchers who were familiar with the general subject matter of the content, each of whom recommended between three and five keywords per paper. Reviewers were told that the papers were all related to the overall topic of COVID-19, so this particular term and its alternative synonymous terms (e.g. "SARS-COV2", "Coronavirus") should not be deemed keywords. They were instructed to nominate keywords freely otherwise, with no prior examples given, choosing words they thought would be useful if they themselves were searching for papers with their perceived topic coverage of the current paper under consideration. They were not instructed to attempt to limit the range and variety of words recommended, nor to adopt their preferred alternative words which they perceived to be equivalent to a perceived keyword occurring in the text.

For each corpus, a list was also constructed of publication source supplied keywords from every paper for which they were available (either explicitly specified in the PubMed keywords field and/or in the Full Text of the papers under a keywords heading). As these keywords are also intrinsically human-generated, they might be expected to provide a means of assessing consistency versus diversity for the human-based approach, when compared with the reviewer derived keywords. The resulting recommended sets of keywords were collated for each corpus and then cleaned by resolving prefix or suffix variants, and terminology differences and synonymity issues by consensus between two authors of this paper. This step reduced the number of distinct keywords to approximately 80% of the original raw keywords. From these cleaned sets of keywords, a simplified framework approach for synthesis was applied by grouping keywords in semantically related groups around dominant concepts, for each identified concept for which there were 3 or more associated keyword occurrences, by the same consensus process as above. This produced ranked sets of 'concepts' which were approximately 10% of the size of the originating cleaned keyword sets. In cases where keywords were phrases rather than single words, it was accepted that they may appear in more than one concept.

2.2. Computation-based method

The computational arm of the work relied on use of text processing tools based on various existing NLP techniques including Term Frequency Analysis and Automated Keyword Extraction [9]. The details of these two different types of approaches will be described below. This software was able to exclude occurrences of stop words (e.g. "a", "the") from the analysis, using a commonly accepted standard set. As all papers deal with COVID-19, this word and equivalent words (e.g. "SARS-COV2", "Coronavirus") were also excluded. In addition, words were removed if deemed to be inappropriate for topics, including general (i.e. non-clinical) abbreviations, numbers, times, dates and geographical locations. All computational analysis was undertaken on the full text of the papers in each corpus, as returned from the original PubMed search for that corpus, including title, author names and affiliations, abstract, main body, references and any other incidental text content from keywords, acknowledgements, illustrations, tables, etc.

For Term Frequency Analysis, frequencies of occurrence of all words were first computed for each document. A token separation and counting algorithm and software from the Natural Language Tool Kit (NLTK) at www.nltk.org was used to perform this task, incorporating a standard set of stop words for English texts. Weighted frequencies were then calculated by using the well-established Term Frequency – Inverted Document Frequency (TF-IDF) relevance method [13] across each corpus of documents. Singular and plural forms of words were combined, as were those with common suffix variants (e.g. adjective and adverb forms).

For Automated Keyword Extraction, each of the three specified defined corpora was treated as a target set of papers for which keywords were to be extracted, and the "General Information" corpus was used as a reference set of papers, with topics anticipated to be mostly different from those for the target set. This reference set was essential for training the logic in the keyword algorithm, by using it to provide counterexamples (i.e. negative cases for potential keywords). It also provided a convenient source for identifying commonly occurring words in a more general context, which could be used later for rationalisation of generated candidate sub-topics. Keywords were generated using the popular software package AntConc https://antconc.en.lo4d.com [14]. This software permitted various configuration and parameter choices, for which we used vendor-supplied default values, and allowed the specification of unacceptable candidate keywords and stop words. The automated keyword analysis was conducted based only on the concatenated texts of respectively the target and reference sets, with no information being returned from the software on term occurrence per document. This software provided a Significance index as a measure of the strength of generated keywords according to its model of keyness, based on prominence of the words relative to the overall bodies of text data in the target and reference sets.

3. Results

Results for both methodological arms are presented here. Note that the contents of the corpora for which results are reported are identical in both arms.

3.1. Human-based results

Table 1 gives details of corpora used and human-based keywords, obtained from sources and reviewers.

Table 1: Corpus and human keyword details

Corpus	*# Papers*	*# Keywords (Supplied)*	*# Keywords (Reviewers)*
Residential Aged Care	50	65	247
Older People >=65 years	59	53	234
Isolation	71	81	263
General Information	70	125	253

The first data column shows the number of papers included in each corpus, which were subjected to further analysis. The second data column shows the number of keywords obtained from the source supplied keyword sets (after cleaning). These values were not strongly correlated with the number of papers because they are indicative of the diversity of content in a corpus, and also variations between different supply sources (e.g. authors, editors). The third data column shows the number of unique keywords nominated by all reviewers for the corpus (after cleaning). It can be seen that these counts are much higher than the supplied keyword counts, due to variety of perceptions and freedom of choice available to the reviewers. These keyword counts are again not strongly correlated with the number of papers in their corpus.

Table 2 shows the lists of the top ranked keywords of both types, in the first two data columns. The keywords are listed in order of frequency, or alphabetically if of the same frequency. The keywords shown in bold are those which are repeated (albeit in variant or composite forms) in both sets for each corpus, being around 50% of the total. The third data column shows the dominant concepts derived by expert consensus from these keywords.

The concept grouping process led to formation of eleven concept groups across all corpora, as follows:

- Aged/Elderly/Older;
- Disease/Cancer/Respiratory;
- Facility/Nursing Home;
- Family/Social;
- Guideline/Treatment;
- Healthcare/Delivery;
- Infection/Transmission/Control;
- Isolation/Quarantine/Prevention;
- Long Term/Residential;
- Pandemic/Epidemic;
- Public Health/Guideline.

While these groupings are intrinsically subjective, they capture a number of related topics which might otherwise have been underserved by considering keywords independently. So for human-based results, an ontological or taxonomic approach may provide better utility. That could be accomplished as here by consensus on common terms, or by constructing a unified preferred vocabulary in advance for selection.

Table 2: Top human keywords and concepts

Corpus	*Top Keywords (supplied)*	*Top Keywords (reviewers)*	*Top Ranked Concepts*
Residential Aged Care	**Care** **Long Term** Pandemic **Infection** **Elder**	**Long Term** **Healthcare** **Infection Control** Transmission **Aged Care** Facility Guideline Workforce	Healthcare/Delivery Long Term/Residential Facility/Nursing Home Pandemic/Epidemic Infection/Transmission/Control Aged/Elderly/Older
Older People (>=65 years)	**Older** **Elder** Care Pandemic Pneumonia Clinical Infection Patient	**Elderly** **Older** People Population Social Isolation Mental	Aged/Elderly/Older Healthcare/Delivery Pandemic/Epidemic Disease/Cancer/Respiratory
Isolation	Health **Social** **Infection** **Isolation** Pandemic Care Mental Exercise	**Isolation** **Social** Case Mental Loneliness Quarantine Epidemiology **Infection Control**	Isolation/Quarantine/Prevention Family/Social Healthcare/Delivery Infection/Transmission/Control Isolation/Quarantine/Prevention Pandemic/Epidemic
General Information	**Infection** **Pandemic** Outbreak **Control** Disease **Prevention** **Quarantine** **Public Health** **Guideline**	**Infection Control** Transmission Containment **Public Health** PPE **Prevention** **Preparedness** Modelling **Epidemic** **Treatment** **Guideline** **Quarantine**	Infection/Transmission/Control Pandemic/Epidemic Disease/Cancer/Respiratory Isolation/Quarantine/Prevention Guideline/Treatment Public Health/Guideline

3.2. Computation-based results

The generation of computational results described in this section can have many variants, from the main configuration parameters for the associated analysis:

- Exclusion of words from consideration if deemed inappropriate as topic candidates;
- Aggregation of word variants including suffixes, vocabulary variations and synonyms;
- Selection of the prominence ranking measures and any related normalisation.

In the results presented here, cleaning was performed as described previously. Choice of measures for rankings are identified below: no optimal measures have yet been claimed for computational topic analysis as theoretical aspects are not well developed.

Table 3: Top computation keywords and concepts

Corpus	*Top Ranked Words (TFA)*	*Top Ranked Words (AKE)*	*Top Ranked Concepts (consensus)*
Residential Aged Care	**care** patient health **resident** **facility** disease **nursing** infection	**care** **resident** **nursing** **facility** home term shelter long	Healthcare/Delivery Long Term/Residential Facility/Nursing Home Pandemic/Epidemic Infection/Transmission/Control Aged/Elderly/Older
Older People (>=65 years)	patient **older** **care** health **adult** disease **social** **elderly** clinical infection	**older** adult **elderly** **geriatric** **social** age **care**	Aged/Elderly/Older Healthcare/Delivery Pandemic/Epidemic Disease/Cancer/Respiratory
Isolation	health **social** patient **isolation** **older** app people **mental**	**isolation** **older** **social** self loneliness **mental** exercise physical	Isolation/Quarantine/Prevention Family/Social Healthcare/Delivery Infection/Transmission/Control Isolation/Quarantine/Prevention Pandemic/Epidemic
General Information	health patient case disease infection transmission cancer public respiratory outbreak control		Infection/Transmission/Control Pandemic/Epidemic Disease/Cancer/Respiratory Isolation/Quarantine/Prevention Guideline/Treatment Public Health/Guideline

Table 3 summarises the computational results using a similar structure to Table 2. The first data column shows the highest ranked words (after cleaning), ordered according to the Term Frequency Analysis measure, TF-IDF. Words occurring in less than 10% of papers in the corpus were excluded, on the basis that those papers were not sufficiently representative of the corpus as a whole. The second data column shows the highest ranked words (after cleaning), ordered according to Automated Keyword Extraction measure, Significance. No corpus analysis is possible for the General Information corpus in this case, because it cannot be both the target and reference set. As before, those words occurring repeatedly in both high-ranking lists are shown in bold, and can again be seen to be around 50%. No separate consensus concept grouping step was undertaken for the computational results, but the Top Ranked Concepts from Table 2 are repeated here in the final data column for ease of comparison. There is again good agreement between these results and this previously established set of concepts.

The weighted average TF-IDF measure for TFA keywords (normalised by the number of papers in which it occurs at least once) was preferred due to equal frequency values being scaled in accord with the number of papers of occurrence. The Significance measure for AKE was found to be a stable statistically based metric not strongly affected by choice of abstract or full text, and consistent with the underlying computational model used to compute keywords.

Figure 1 provides graphical presentations of these two measures, for the Residential Aged Care corpus. Both graphs present as smooth curves with similar profiles, as the sets of keywords have been listed in descending order of the measure values. After the steep decrease for the first few words, there is knee point separating a fairly constant and more gradual decrease in the measure value over many successive words. Finally there is a very long and flat tail containing all the remaining candidate words.

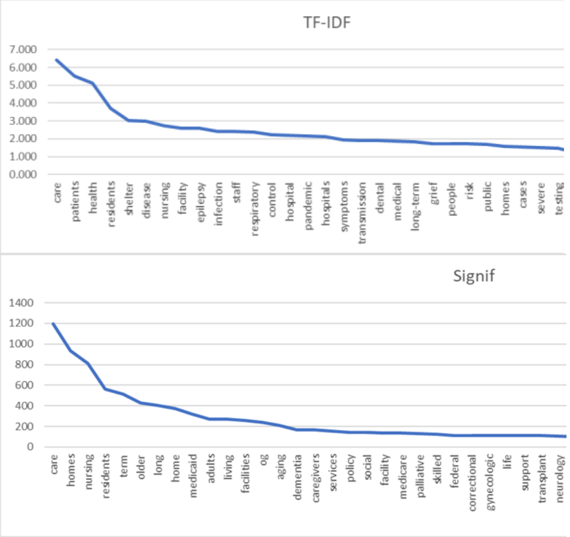

Figure 1: Comparison of computational rankings

4. Conclusion

The work presented here, using a simple word or keyword model for identifying topics, was prompted by the small size and restricted scope of the corpora of interest. More sophisticated context based or machine learning driven models would not be feasible in this setting, as there would be insufficient data to assure their convergence. Recent work has suggested emerging statistical approaches for such short text situations [5] but has not provided a comparison with traditional methods or human performance. In our case, both human and computational approaches yielded similar but not identical lists of prominent words as candidate topics. These words were intuitively relevant for describing topics but were in a sense “obvious” i.e. not requiring any specialised or detailed knowledge of the domain covered by the text.

As more words of lower prominence are considered in this case, the similarity relationship between human and computational approaches deteriorates and the rankings become disrupted. This is due mostly to the imprecision of our chosen measures for small corpora such as these, and the lack of influence of context. It could also be argued that a number of the words designated as topics or concepts, are widely used in many areas of healthcare studies and not peculiar to COVID-19 (e.g. health, care, patient, hospital). Determining which words are “generic” in this way, so that they can be excluded or their prominence can be downgraded in the measures, presents a difficult problem because a very large body of well-selected text would be needed to derive this information. We have not attempted to address this issue in this paper, but it would form a natural focus for improvement of results, in future extension of the work.

The computational results also indicate a need for further work on inferring semantic value from text, which might provide better topic descriptions than the current simple approach. Multi-word or phrase-based analysis (e.g. using a sequence of adjacent words) would enable context to be incorporated in the identification of candidate topics.

A major limitation of the study was that the amount of text available was small by comparison with the diversity of text content. The original specific corpora were very broad in coverage, and papers associated with them covered a wider range of topics and at lower frequency in the text, than the analysis algorithms were deigned to address. Furthermore, the reference corpus was effectively a superset of the topics in any of the other corpora, so the training on counterexamples was not as powerful as expected.

An interesting question is whether, in cases with a very much larger number of papers per corpus, similar performance could be obtained by using only the title, keywords and abstract content rather than full papers. This is plausible on the basis that these components are deliberately constructed to provide the most important information pertaining to the paper, while the main body of text contains looser explanatory content and the language used is often strongly stylised according to the preferences of the authors. This is another aspect of our work which deserves further investigation.

In conclusion, it appears from this study that it is possible for computational approaches to topic analysis based on term frequencies and keyword extraction to be able to perform comparably with human experts in identifying topics in limited corpora settings, provided some latitude of concept association is included. The results do not support either of the computational approaches as being superior to the other, as they differ from the human performance results in contrasting aspects. Topics associated with highly repeated terms which had a stronger tendency towards being specific were identified by term frequency, while dominant concept groups associated with more generic topics emerged more strongly in the keyword extraction case.

Such approaches may therefore be of value in supporting ongoing evidence retrieval for knowledge-based projects such as ELDAC, but may yield best results if used in combination, as demonstrated here. This ameliorates the restriction on using more advanced computational approaches based on deeper artificial intelligence and machine learning techniques, which require very large datasets in order to train adequately for the embedded contextual patterns.

5. Acknowledgements

This work formed part of the ELDAC project funded by the Australian Government Department of Health. The assistance of the following volunteers from Flinders University in the human-based arm of the project is gratefully acknowledged: Nicky Baker, Chris Barr, Didy Button, Stacey George, Claire Gough, Arun Gunachandran, Anna Hall, Judith Hocking, Susan Rampling, Maayken van den Berg, Sally Vuu.

6. References

[1] M. J. Grant and A. Booth, “A typology of reviews: an analysis of 14 review types and associated methodologies”, Health Information and Libraries Journal 26, 2009, pp. 91-108.

[2] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera, “Systematic review automation technologies”, Systematic Reviews 3(1), 2014, p.74.

[3] H. Li and K. Yamanishi, “Topic analysis using a finite mixture model”, Information Processing & Management 39(4), 2003, pp. 521-541.

[4] R. Alghamdi and K. Alfalqi, “A survey of topic modeling in text mining”, International Journal of Advanced Computer Science Applications 6(1), 2015, pp. 147-153.

[5] J. Qiang, Z. Qian, Y. Li, Y. Yuan and X. Wu, “Short text topic modeling techniques, applications, and performance: a survey”, IEEE Transactions on Knowledge and Data Engineering, 2020 (in press).

[6] C. Gabrielatos “Keyness analysis: Nature, metrics and techniques”, in: Corpus approaches to discourse: A critical review, C. Taylor and A. Marchi, Eds. New York, NY: Routledge, 2018, pp. 225–258.

[7] N. Firoozeh, A Nazarenko, F. Alizon, and B. Daille, “Keyword extraction: issues and methods”, Natural Language Engineering 26(3), 2020, pp. 259-291.

[8] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, “Using text mining techniques for extracting information from research articles”, in Intelligent Natural Language Processing: Trends and Applications, K. Shaalan et al., Eds. Berlin: Springer, 2018, pp. 373-397.

[9] S. K. Bharti, K. S. Babu, and S. K. Jena, “Automatic keyword extraction for text summarization: a survey”, arXiv preprint arXiv:1704.03242, 2017, 12 pp.

[10] O. Medelyan, E. Frank, and I. H. Witten, “Human-competitive tagging using automatic keyphrase extraction”, Proceedings 2009 Conference on Empirical Methods in Natural Language Processing, Singapore 6-7 August 2009, pp. 1318-1327.

[11] R. A. Damarell, N. May, S. Hammond, R. M. Sladek, and J. J. Tieman, “Topic search filters: a systematic scoping review”, Health Information & Libraries Journal 36(1), 2019, pp. 4-40.

[12] J. Kaur and V. Gupta, “Effective approaches for extraction of keywords”, International Journal of Computer Science Issues 7(6), 2010, pp. 144-148.

[13] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, “Interpreting TF-IDF term weights as making relevance decisions”, ACM Transactions on Information Systems 26(3), 2008, pp. 1-37.

[14] L. Anthony, “Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom”, IEEE Transactions on Professional Communication 49(3), 2006, pp. 275-286.