Improving MeSH Classification of Biomedical Articles using Citation Contexts

Bader Aljaber[a], David Martinez[a,b,*], Nicola Stokes[c], James Bailey[a,b]

[a]*Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia*
[b]*NICTA, Victoria Research Laboratory, The University of Melbourne, Victoria 3010, Australia*
[c]*School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland*

[*]Corresponding author. Fax: +61 3 9348 1184
Email address: davidm@csse.unimelb.edu.au (David Martinez)

Preprint submitted to Journal of Biomedical Informatics *May 4, 2011*

Abstract

Medical **S**ubject **H**eadings (MeSH) are used to index the majority of databases generated by the National Library of Medicine. Essentially, MeSH terms are designed to make information, such as scientific articles, more retrievable and assessable to users of systems such as PubMed. This paper proposes a novel method for automating the assignment of biomedical publications with MeSH terms that takes advantage of citation references to these publications. Our findings show that analysing the citation references that point to a document can provide a useful source of terms that are not present in the document. The use of these citation contexts, as they are known, can thus help to provide a richer document feature representation, which in turn can help improve text mining and information retrieval applications, in our case MeSH term classification. In this paper, we also explore new methods of selecting and utilising citation contexts. In particular, we assess the effect of weighting the importance of citation terms (found in the citation contexts) according to two aspects: i) the *section* of the paper they appear in, and ii) their *distance* to the citation marker.

We conduct *intrinsic* and *extrinsic* evaluations of citation term quality. For the intrinsic evaluation, we rely on the UMLS Metathesaurus conceptual database to explore the semantic characteristics of the mined citation terms. We also analyse the "informativeness" of these terms using a class-entropy measure. For the extrinsic evaluation, we run a series of automatic document classification experiments over MeSH terms. Our experimental evaluation shows that citation contexts contain terms that are related to the original document, and that the integration of this knowledge results in better classification performance compared to two state-of-the-art MeSH classification systems: MeSHUP and MTI. Our experiments also demonstrate that the consideration of *Section* and *Distance* factors can lead to statistically significant improvements in citation feature quality, thus opening the way for better document feature representation in other biomedical text processing applications.

Keywords: Citation contexts, document expansion, biomedical text classification, MeSH

terms

1. Introduction

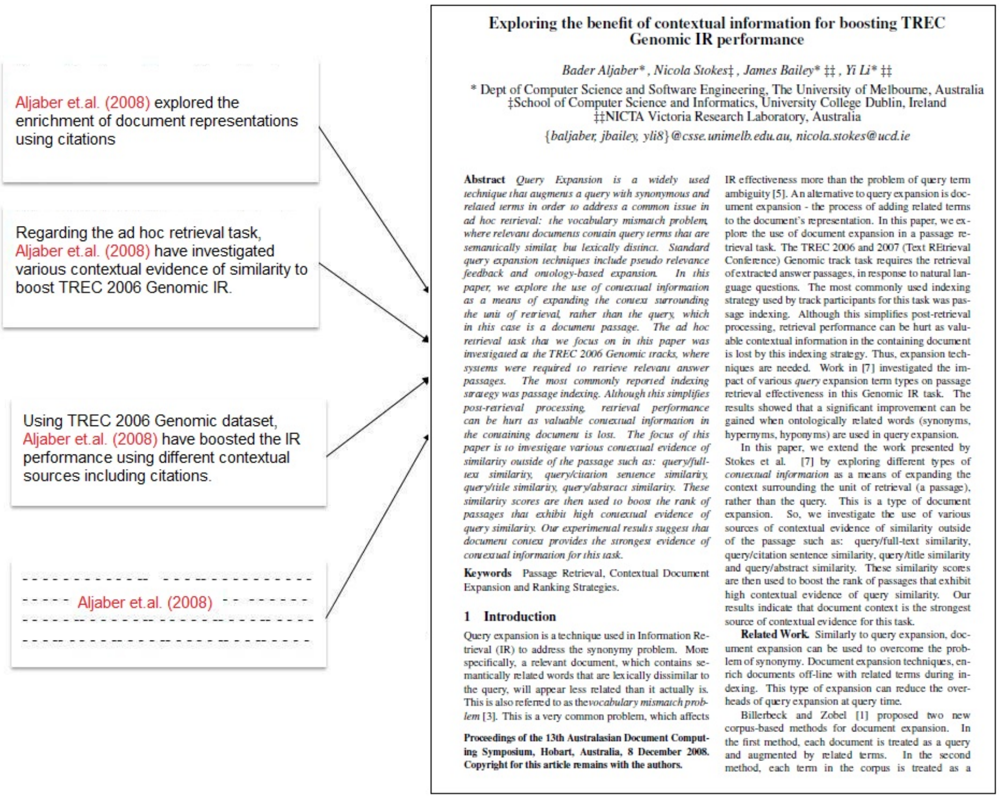
Citations are extensively used in academic publications in order to refer to related work, or to point to extra information complementing what is being said. An example of a citation is shown in Figure 1. Each citation provides a link to the reference material and a context that describes some aspect of it. A citation context is the text surrounding citation markers used to refer to other publications. These text snippets can be a useful source of terms, such as relevant synonyms and related vocabulary that is not present in the document. For instance, the term "enrichment" that is used in one of the citations does not occur at all in the cited document, which refers to this concept with the term "expansion". The use of these citations can therefore help to provide a richer document feature representation. Previous work has identified the usefulness of this source of information for applications such as Text Mining [1, 2, 3], and Information Retrieval (IR) [4, 5].

In recent times, text analysis applications have been the object of extensive study, specially in areas such as biomedicine where there has been a huge growth in the amount of information published. In the biomedical domain alone, around 1,800 new papers are published daily [6]. As of September 2009, MEDLINE, which is the largest collection of bibliographic records on the biomedical literature, contained more than 19 million references, and it is estimated that the employees of the National Library of Medicine[1] (NLM) add between 1,500 and 3,500 new references to the database every day [7]. In order to make these publications more accessible, MeSH[2] (***Me***dical ***S***ubject ***H***eading) terms are used to index all these entries; a time consuming process which could significantly benefit from an automatic text classification solution.

Traditionally, text processing techniques represent documents by using the publication's original source text, which consists of features such as terms and phrases. Moreover, many tools, such as text classifiers [8, 9, 10], use the bag-of-words (BOW) model to represent the

[1] http://www.nlm.nih.gov

[2] http://www.nlm.nih.gov/mesh/



Aljaber et.al. (2008) explored the enrichment of document representations using citations

Regarding the ad hoc retrieval task, Aljaber et.al. (2008) have investigated various contextual evidence of similarity to boost TREC 2006 Genomic IR.

Using TREC 2006 Genomic dataset, Aljaber et.al. (2008) have boosted the IR performance using different contextual sources including citations.

Aljaber et.al. (2008)

Exploring the benefit of contextual information for boosting TREC Genomic IR performance

*Bader Aljaber** , *Nicola Stokes‡* , *James Bailey** ‡‡ , *Yi Li** ‡‡

* Dept of Computer Science and Software Engineering, The University of Melbourne, Australia
‡School of Computer Science and Informatics, University College Dublin, Ireland
‡‡NICTA Victoria Research Laboratory, Australia

{*baljaber, jbailey, yli8*}*@csse.unimelb.edu.au, nicola.stokes@ucd.ie*

Abstract *Query Expansion is a widely used technique that augments a query with synonymous and related terms in order to address a common issue in ad hoc retrieval: the vocabulary mismatch problem, where relevant documents contain query terms that are semantically similar, but lexically distinct. Standard query expansion techniques include pseudo relevance feedback and ontology-based expansion. In this paper, we explore the use of contextual information as a means of expanding the context surrounding the unit of retrieval, rather than the query, which in this case is a document passage. The ad hoc retrieval task that we focus on in this paper was investigated at the TREC 2006 Genomic tracks, where systems were required to retrieve relevant answer passages. The most commonly reported indexing strategy was passage indexing. Although this simplifies post-retrieval processing, retrieval performance can be hurt as valuable contextual information in the containing document is lost. The focus of this paper is to investigate various contextual evidence of similarity outside of the passage such as: query/full-text similarity, query/citation sentence similarity, query/title similarity, query/abstract similarity. These similarity scores are then used to boost the rank of passages that exhibit high contextual evidence of query similarity. Our experimental results suggest that document context provides the strongest evidence of contextual information for this task.*

Keywords Passage Retrieval, Contextual Document Expansion and Ranking Strategies.

1 Introduction

Query expansion is a technique used in Information Retrieval (IR) to address the synonymy problem. More specifically, a relevant document, which contains semantically related words that are lexically dissimilar to the query, will appear less related than it actually is. This is also referred to as the*vocabulary mismatch problem* [3]. This is a very common problem, which affects IR effectiveness more than the problem of query term ambiguity [5]. An alternative to query expansion is document expansion - the process of adding related terms to the document's representation. In this paper, we explore the use of document expansion in a passage retrieval task. The TREC 2006 and 2007 (Text REtrieval Conference) Genomic track task requires the retrieval ofextracted answer passages, in response to natural language questions. The most commonly used indexing strategy used by track participants for this task was passage indexing. Although this simplifies post-retrieval processing, retrieval performance can be hurt as valuable contextual information in the containing document is lost by this indexing strategy. Thus, expansion techniques are needed. Work in [7] investigated the impact of various *query* expansion term types on passage retrieval effectiveness in this Genomic IR task. The results showed that a significant improvement can be gained when ontologically related words (synonyms, hypernyms, hyponyms) are used in query expansion.

In this paper, we extend the work presented by Stokes et al. [7] by exploring different types of *contextual information* as a means of expanding the context surrounding the unit of retrieval (a passage), rather than the query. This is a type of document expansion. So, we investigate the use of various sources of contextual evidence of similarity outside of the passage such as: query/full-text similarity, query/citation sentence similarity, query/title similarity and query/abstract similarity. These similarity scores are then used to boost the rank of passages that exhibit high contextual evidence of query similarity. Our results indicate that document context is the strongest source of contextual evidence for this task.

Related Work. Similarly to query expansion, document expansion can be used to overcome the problem of synonymy. Document expansion techniques, enrich documents off-line with related terms during indexing. This type of expansion can reduce the overheads of query expansion at query time.

Billerbeck and Zobel [1] proposed two new corpus-based methods for document expansion. In the first method, each document is treated as a query and augmented by related terms. In the second method, each term in the corpus is treated as a

Proceedings of the 13th Australasian Document Computing Symposium, Hobart, Australia, 8 December 2008. Copyright for this article remains with the authors.

Figure 1: An example of a document being cited.

documents in which each feature corresponds to a single word. The BOW model in Natural Language Processing (NLP) and IR is a popular method for representing documents, as it is very simple and highly effective. However, this representation ignores semantic relationships between terms. Hence, the selection and weighting of features must be carefully done.

This paper examines different ways of enriching the feature representation by relying on external resources such as the text surrounding citations of a scientific publication (i.e., citation contexts), and the conceptual relations found in the Unified Medical Language System (UMLS) Metathesaurus[3]. The main idea is to explore ways to better extend the representation of a given document by the terms that are used to refer to it. Looking at Figure 1, all the text snippets (citation contexts) citing that document are used to enrich the representation of that document. We also present an analysis of the types of terms that are

[3] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

found in citation contexts, and propose a way to obtain the most benefit from these types of features in a MeSH term classification task. We explore whether citation contexts are a useful alternative source of semantically related terms, which can be used to strengthen the *topical focus* of a document's original feature representation. However, these features need special consideration - in particular with respect to selection and weighting - in order to achieve an improvement over baseline performance. These are the main questions that we address in this work:

1. What kind of relationships exist between citation terms and the full-text content of documents? We analyse and identify the type of terms that are acquired from citations to better understand their contribution, and also to learn if citation contexts contain both lexically equivalent terms and many related terms such as synonyms, near-synonyms and spelling variants.
2. Does document layout information have an impact on the usefulness of those terms? In other words, are certain sections of a paper more likely to contain useful citation terms? We investigate weighting the citation terms based on the sections containing them.
3. Can the distance (in words) of the citation terms to the citation marker influence the usefulness of those terms? We investigate weighting the citation terms based on the distance between them and their citation markers.
4. To what extent can the length of citation contexts affect their usefulness? The citation context is extracted based on a window size parameter. The window size is the number of extracted terms before and after the citation marker.

By exploring the above questions, we test our hypothesis that these citations characteristics can be used to optimising a text classification model for scientific publications. We evaluate this hypothesis by evaluating our new model in the context of MeSH classification task with respect to two state-of-the-art systems. There are two main novel contributions of the work presented in this paper. ***First***, we provide a novel *intrinsic* evaluation methodology for determining the quality of citation terms (cf. Section 4) by analysing i) the semantic characteristics of the citation terms (e.g. whether they are synonyms/hypernyms) and ii) the relationships between the following factors:

- The presence of synonyms/hypernyms with respect to the document section in which they occur.
- Citation term entropy (or informativeness) and the document sections where these terms occur in.
- Citation term entropy and the distance to their citation markers.

Second, we evaluate the citation terms *extrinsically*, where the objective is to see if our observations on citation quality result in better document representation, and hence more accurate text classification of biomedical publications (more details will be given in Section 7). We use the terms in the citations to improve document classification, and analyse the effect of the following parameters: (i) section (and subsections) of the paper where the citation comes from, (ii) distance of the term to the citation marker, (iii) citation context window size, and (iv) type of terms (synonyms, hypernyms) in the citation. Hence, we focus in our experiments on feature engineering, and specifically on how best to select and weight these features. We also compare our approach to two state-of-the-art MeSH tag classification systems, namely MTI [11] and MeSHUP [12]. To the best of our knowledge, this is the first published application of citation contexts in a MeSH classification task.

The remainder of the paper is organised as follows. In Section 2 we discuss related work. We then introduce the dataset and resources used in our experiments in Section 3. The intrinsic evaluation over our dataset is presented in Section 4. We then move on to the text classification task, and describe our document representation, experimental setting, results, and findings in Sections 5, 6, 7, and 8 respectively. Finally, we present our conclusions and future work in Section 9.

2. Related work

In this section, we provide an overview on work that analyses citation contexts, and we explore how these have been applied to language technology applications. We then discuss the relationship between citation contexts and anchor text, which has been successfully applied by the IR community in the area of Web search. Finally, we describe related work on the text classification task, which we will use for extrinsic evaluation.

2.1. Analysis of citation contexts

Citations and their use have been of great interest to researchers. One of the earliest studies on the importance of citations for analysis of scientific literature was published by Garfield in [13]. In more recent work, the study of the text surrounding citations (also referred as citation sentences or *citances* [1]) has been used to determine the relationship between the two papers connected by that citation, defining a *citation function* [14, 15].

Related work by Teufel and Moens [16, 17], and Nanba et al. [18, 19, 20] automatically analyses citation contexts. Teufel and Moens develop an argumentative zoning [4] technique, which is a discourse classification technique that labels sentences according to their role in the authors' argument, e.g. contrasting, basis, and background. Their method can identify the novel claim or contribution of a cited paper by analysing its citations. This classification technique is used to generate summaries of the cited papers by showing sentences that support the specific rhetorical role. Their most recent work has shown that the approach can be applied to fine-grained analysis and different domains with high annotator agreement. Nanba et.al. published some interesting work that explores characteristics of citations; they analyse citations of research papers and automatically classify citation links based on their motivations into three categories, using cue phrases and 160 rules. The three categories are (i) a comparison to other related papers (either negatively or positively) (ii) building on other related work (iii) others that do not fall into either of the previous two classes. This categorization scheme is used to build a system for reviewing and surveying academic literature.

[4] Argumentative Zoning [21]; http://www.cl.cam.ac.uk/ sht25/az.html

Another approach to analyse citation contexts is to study the terms found in them. Ritchie, Teufel & Robertson [22] identified the words from around the citations that specifically referred to the cited paper, both manually and automatically (using a fixed window size). They found that there was overlap between the citing terms and important terms in the original document. Also, combining citing terms with terms in the original document (using the *tf-idf* weighting scheme) was found to be useful for ranking relevant terms to represent a document.

2.2. Applications of citation contexts

Regarding more specific applications of citation contexts, early work by Nakov et al. [1] focuses on the utility of citations for managing life science literature. They identify a number of promising applications of citations in this domain: as a source of unannotated comparable corpora, summarisation of the target papers, synonym identification and disambiguation, entity recognition, relation extraction, and improved citation indexes for document retrieval. In the same article, Nakov et.al. also introduce the idea of using citation contexts as comparable corpora for automatic paraphrase extraction. These citation contexts have been used to support automatic paraphrasing. Thus, the extracted paraphrases have to cite the same target article. In particular, the authors propose a paraphrase extraction algorithm that identifies the relationship between two named entities; such as genes, proteins or MeSH terms[5], such as Neuregulins and Brain-Derived Neurotrophic Factor. In summary, named entities found in each citation sentence are identified; then, based on a dependency parser, the path between them is extracted and a paraphrase built. Finally, the candidates of name entities are ranked to select only those above a given threshold.

Another possible application of citation contexts is automatic summarisation. Mohammad et. al. [23] propose a method that produces an automatically generated multi-document survey. The method is built on four summarisation systems that use citation terms. Com-

[5] MeSH stands for ***Me***dical ***S***ubject ***H***eadings which are part of a large controlled vocabulary of topic terms used for indexing journal articles and books in the Life Sciences arena, and managed by the United States National Library of Medicine (NLM); http://www.nlm.nih.gov.

pared with summarisation based on full-text document, citation terms provide additional information, which cannot be found elsewhere. Elkiss et.al. [3] provided a quantitative analysis of the benefits of citation contexts with regards to similar applications, such as summarisation and information retrieval. In particular, they examined the relationship between the abstract and citation contexts of a given scientific paper. Their experiments show that citation contexts tend to have extra focused information that is not present in the abstract. Therefore, they suggest that citation contexts can be utilized as a different kind of supplementary summary to the traditional abstract.

Also for summarisation of information, a research tool called the Citation-Sensitive In-Browser Summariser (CSIBS) was introduced by Wan et.al. [24, 25]. When researchers read the academic literature, to enhance their knowledge and explore new topics and methodologies, they come across citations to other related works. To save time in deciding whether the cited work is worth reading or not, a research tool to help manage the literature browsing task was built. The inventors of CSIBS conducted a user requirements analysis [26] for researchers (especially, in the biomedical field) while they browsed through the academic literature. They found that they often lacked the necessary contextual information for interpreting the interestingness of the citations they encountered. Thus, CSIBS was built to provide researchers with a summary of the cited document. CSIBS can be used as a web service attached to an existing publication repository. A qualitative evaluation showed that the generated summaries provide useful information that was sufficient for judging the relevance of cited documents [26, 25].

A straightforward application of citation contexts, and the one we will explore in this paper, is text classification. In previous work, citation terms have been used mostly for document expansion. That is, the document representation of a publication (usually BOW) is augmented with terms found in sentences surrounding citations of the paper in the rest of the document corpus [4, 27, 28]. In our previous work, published in [29], we investigated the usefulness of citation terms in a document clustering task. Our results indicated that citation terms are, in general, useful when combined with the original representation. Also, we investigated citation terms based on different levels of topic granularity and found that citation terms tend to capture general topic keywords rather than specific ones. However, the citation terms can introduce noise if they are not related to the general topic of the cited paper. In our present work, we analyse the relationships between terms in the original document and citation terms in order to define a better model. We extend our previous work by also investigating factors that affect the usefulness of the citation terms in a different text processing task - supervised document classification.

Citation contexts have been also applied to information retrieval (IR). Bradshaw [28, 27] introduced a novel automatic document indexing scheme based on citations, called Reference Directed Indexing (RDI). RDI uses terms in citation sentences to index a cited article. Documents are then ranked with respect to the following metrics: the relevance score between document index terms (from the citation sentences) and the query terms, and the number of papers citing that document. Hence, highly cited documents will be ranked higher than documents with lower numbers of citations even if their term indexes have the same number of query terms. The performance of RDI was evaluated against the standard vector-space model, which uses the *tf-idf* weighting method and the Cosine similarity metric. RDI achieved better precision on the top 10 retrieved documents (statistically significant at 99.5% confidence) [30, 31, 27].

In a more recent work [5], Ritchie et.al. presented the results of experiments using terms from citations for scientific literature search. For every document, they combined terms from the full-text document itself and terms used by other authors to refer to that document. The influence of weighting citation terms differently relative to document terms was measured. A set of weights was used to evaluate the citation terms. As a result, the IR performance is improved when citation terms are weighted more. Also, they used a range of standard performance measures and t-test for statistical significance and ran the queries through several standard retrieval models, as implemented in the Lemur Toolkit[6]: Okapi BM25, KL-divergence and Cosine similarity. In each run, 100 documents were retrieved per query. Overall, the IR performance is increased with citation terms, for all models, for all

[6] http://www.lemurproject.org/

measures, with the exception of Okapi run [5].

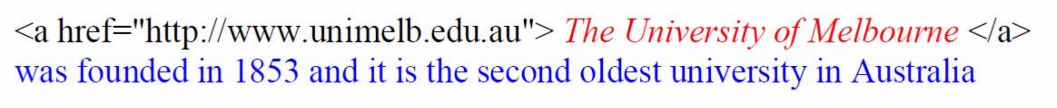
Ritchie et.al. in [4] compare different lengths of citation contexts for IR, including: no context, the entire citing paper, different fixed window sizes, and sentence boundaries. The results show that adding citation terms to the full-text representation can improve the performance of information retrieval systems at different levels. More specifically, longer citation contexts (but not the whole citing documents) tend to be better. The authors conclude that applying natural language processing techniques to identify the related citation terms can bring further improvement.

Our work is related to [5, 4, 3], who used citation terms with original full-text to boost systems such as IR and text summarisation. The main differences of our approach are our application task (text classification), the implementation of intrinsic evaluation, and the reliance on sophisticated term-weighting models based on a variety of parameters: sections that the terms come from, distance to the citation markers, semantic relationships from a knowledge-base, and window size.

Finally, a recent body of work has focused on context-aware citation recommendation. Sugiyama et.al. [32] presented a supervised classification system that takes a draft (unpublished) paper as input and decides whether there are sentences in that paper which need citations. They conducted their experiments over two supervised classifiers, namely maximum entropy (ME) and support vector machines (SVM). Also, they extracted different kinds of features such as unigrams, bigrams, proper nouns, and previous and next sentence. The results showed high accuracy scores (0.882) when proper noun and previous and next sentence features are used. Another related citation recommendation system has also been proposed by He et.al. [33]. They implement a prototype system in CiteSeerX, where a citation context and the title and abstract are submitted, and a set of ranked relevant recommendations are retrieved.

2.3. Anchor Text use in Web Retrieval

Anchor text is another way of referring to related information, and consists of a piece of clickable text that links to a target Web page. More precisely, the *anchor text* is defined as the text encompassed by a '<*a href*' tag in an HTML document. For instance, Figure 2 shows an example of a text snippet of an anchor text; where the words '*The University of Melbourne*' represent an anchor text snippet, and the words '*was founded in 1853 and it is the second oldest university in Australia*' represent the extended anchor text.



```
<a href="http://www.unimelb.edu.au"> The University of Melbourne </a>
was founded in 1853 and it is the second oldest university in Australia
```

Figure 2: An example of an anchor text.

Extended anchor text refers to text surrounding the vocabulary outside of the hypertext link, which is defined by a fixed window size. In addition, researchers have included surrounding *headings* and other *highlighted text fragments* in their extended anchor text definition. Therefore, the anchor text and the extended anchor text in web pages are similar to the citation marker and citation context in academic documents. The link structure of the Web, including *anchor text* and *extended anchor text*, has been studied extensively in IR and exploited to advantage in some retrieval tasks [34].

There is a clear parallel between the *anchor text* (or *extended anchor text*) and *citation contexts* of scientific literature: they both provide a semantic linkage between documents. However, there are also a number of critical differences between them: (i) anchor text links in web pages are not always informative, as they may be just commercial or navigational links, whereas links of citation contexts are curated and purposefully inserted; (ii) links of anchor text can link to various types of objects, such as web pages and pictures, whereas links of citation contexts always link to textual documents; (iii) links of anchor text can be changed at any time, whereas links of citation contexts cannot be changed once the paper is published in journals or proceedings; and (iv) the window size of extended anchor text is relatively small compared with the window size of citation contexts.

Many popular literature search engines, such as CiteSeerX[7] [35] and Google Scholar[8],

[7] Scientific Literature Digital Library, http://citeseerx.ist.psu.edu

[8] Google search engine, for peer-reviewed scholarly literature, http://scholar.google.com

also use the links between articles and documents provided by citations to enhance their ranked retrieval results. These retrieval systems provide researchers with a means of crawling and navigating through the network of scholarly scientific articles (that is, the citation graph) in a particular domain. Citation links have also been used in those search engines to analyze research trends, and discover the relationships between publications and their ranking in terms of the number of times they have been cited [36]. There are two well-known algorithms which exploit link structure in this area: *PageRank* which is a query-independent link analysis algorithm [37] and *HITS* which is a query-dependent algorithm and stands for Hyperlink Induced Topic Search [38].

Past research on the TREC Web retrieval tasks was not able to show the effectiveness of anchor text [39]. One of the reasons for this could be that the document collections and link graphs being used were small. However, the TREC 2009 Web Track collection was very large compared with previous collections, and using this data Koolen and Kamps [39] re-examined the importance of anchor text for ad hoc search. They found that at early precision, the use of anchor text even outperformed full-text. With regards to overall precision, they showed that the combination of anchor text and full-text achieved the best result. In this article, the authors also investigated the relationship between the performance and the size of the dataset (original documents and anchors). They observed a clear decrease of the effectiveness of anchor text when the number of anchors was reduced by downsampling. However, when they applied downsampling to the original documents in the collection, they observed that the relative effectiveness of anchor text decreased over the original full text. As a result, Koolen and Kamps (2010) concluded that the use of anchor text is most effective for larger collections.

2.4. Text Classification for the biomedical domain

Finally, we describe related work on text classification for the biomedical domain. There has been interest from many research groups in developing text mining tools [40, 41, 42] for the biomedical domain. Cohen and Hersch [43] provide a survey of work on this area. Some of this work has been centered around the MeSH ontology from the NLM. MeSH terms (classes) are used to manually index all the entries (articles) into MEDLINE, which is the largest collection of bibliographic records of the biomedical literature. These terms are organised into a hierarchy of 24,000 terms, making automation challenging, and the use of automatic aids for the process has been pursued for a long time, as the NLM's Indexing Initiative[9] illustrates. As a result of this initiative the Medical Text Indexer (MTI), based on ngram search, was built by NLM. MTI is a text processing system which relies on semantic relationships to retrieve a ranked list of MeSH terms according to a medical journal, using knowledge from the Unified Medical Language System (UMLS) and information from the MEDLINE database of citations [11].

Most research on automatic MeSH classification does not consider the full set of MeSH tags. Instead, techniques focus on a reduced version of the hierarchy, as is the case in [7], where the categories of MeSH terms (classes) are generalised to the second level of the tree, resulting in a set of 114 classes. For this system, techniques rely on automatic rule generation, and their best performances reach an f-score in the high fifties. Other approaches also decided to focus on a smaller subset of MeSH tags; recent work by the NLM research group Sohn et.al. [44] involved choosing 20 MeSH terms covering different frequency ranges for their experiments. Then Sohn et.al. developed an approach motivated by active learning to construct an "optimal" training set, obtaining an average precision of over 50%, significantly better than the baseline.

The MeSHUP system, which is developed by [12], explores the combination of different Machine Learning (ML) approaches to perform classification over the full class-set. Additionally, they evaluate their results on an IR task, from a ranked output of MeSH terms. The results show that their method is able to improve the performance of MTI, but a limitation of the evaluation is that they only present the results for the optimal cut-off of the ranking.

[9] http://ii.nlm.nih.gov

3. Dataset and knowledge sources

The corpus used in our experiments is a subset of the TREC Genomic 2006/2007 document collection [10], which consists of 162,259 full-text HTML journal articles, published electronically via Highwire Press. This collection is the largest publicly available collection of full-text articles; previous collections consisted of titles, abstracts and keywords only, due to the reluctance of publishers to release pay-per-view content even for academic use. The TREC Genomic collection is also a valuable resource because these full-text documents facilitate the identification and collection of citation contexts from the main body of these publications. Therefore, every document can be represented by two different representations, namely: *original* full-text and *citation* representations. The original full-text representation consists of terms found in the document itself; whereas the citation representation consists of terms found in citation contexts from other documents that refer to the target document.

Identifying the right context for each citation is not an easy task. The relevant text to a marker can be located before or after it, or even both; it can consist of a few words, or go on for many sentences. In this work we rely on a 50-word window at each side of the target word (truncated if there is a paragraph break), an approach that has produced good results in other previous works. For example, Ritchie et.al. in [4] compare different lengths of citation contexts and investigate the effectiveness of those various lengths of citation context around the citation markers, in order to better select good terms in the context of document retrieval task. That range of citation context length includes: no context, the entire citing paper, different fixed window sizes, and sentence boundaries. Their results show that longer citation context length (but not the whole citing documents) is better. Note also that in our work we rely on a BOW representation, and therefore we do not need syntactically valid sentences. Apart from the 50-word windows, we decided to perform an experiment with different window sizes in Section 7, including the full paragraph the marker is in.

With regards to the document collection, we only rely on the subset of documents that has at least one incoming citation in the collection, and that leaves us with 3,475 documents.

[10] http://ir.ohsu.edu/genomics/

We did not perform any sophisticated matching of citations to papers, and we built our dataset based on the explicit references to PubMed-identifiers. This makes us discard some citations, but allows us to experiment on the most explicit, easy-to-parse references. The final collection contains 16,090 citation contexts overall, with an average of 4.63 contexts (the standard deviation is 7.7 contexts) and 33.64 terms for each context.

Each document in the collection has manually-assigned MeSH terms, and this will allow us to experiment on text classification. Our goal will be to automatically predict these tags. As was mentioned earlier in Section 2.4, MeSH terms are manually assigned to all documents in MEDLINE by the NLM, and are organised into a hierarchy of 24,000 terms, making automation challenging.

In our subset of the TREC Genomic 2006/2007 document collection, we have an unbalanced class distribution. There are also some MeSH terms (classes) which have been assigned to only a small number of documents in our dataset. As a result, and like previous work described in Section 2.4, our experiments will rely on a subset of this tagset, by selecting the 20 most frequently occurring MeSH terms in our document collection (see Table 1 for the full list).

Finally, for ontological knowledge, we rely on the Metathesaurus, developed by the NLM, which contains information about biomedical and health related concepts. Its hierarchical structure also captures the relationships between concepts, e.g. 'head trauma' is_a_type_of 'injury'. This will allow us to study the relationships between terms from different sources (original full-text document and citations). We use the UMLS-query Perl module [45] to interface with the Metathesaurus and extract related words. UMLS version 2009AA was used for our experiments.

We focus on two types of relationships between terms:

- **Synonyms (SYN):** Synonyms are distinct lexical forms for identical or very similar meaning concepts. For example, *injury* and *trauma* , or *hemorrhage* and *blood loss.*

- **Hypernym (HYP):** A hypernym is a word whose semantic range includes another word. For example, *injury* is a hypernym of *burn* , and *organism* is a hypernym of *bacteria.*

4. Analysis of citation term characteristics

In this section we conduct an intrinsic analysis of the kinds of terms that we find in citation contexts, and the effect of influential factors, such as the *Sections* they are contained in and the *Distance* to citation markers, on the "quality" of citation terms. For a quantitative analysis of these terms, we rely on two indicators: (i) Metathesaurus, an extensive domain-specific thesaurus that provides links with semantic relationships between different terms, and (ii) Shannon's entropy measurement [46], which estimates the average information content of a message, or in this case a single term. These allow us to intrinsically evaluate terms found in citation contexts, independently of other applications. In previous work in [29], we also developed an approach for intrinsic evaluation, by relying on pairwise similarity between citations and original documents. This method showed that there are substantial differences between them. Our new approach will provide more insight on the types of relationships among the terms from different sources, regions of the paper, and distance to the marker.

Thus, we will first analyse the relationship between the terms in the original full-text document and the citations, by employing the thesaurus. For our second experiment, we will rely on both the thesaurus and entropy measures to analyse the type of citation terms according to two parameters: (i) the *Section* of the paper they occur in, and (ii) the *Distance* to the citation marker.

4.1. Semantic relationships between terms

In this subsection, we rely on the Metathesaurus to study the way in which citation terms and original terms are related. Two factors are measured: (i) the overlap between the original full-text representation and the citation contexts, and (ii) the relationship between novel terms in the citation contexts and the original terms in the full-text representation. A novel (non-overlapping) term in a citation context is a term that occurs in a citation context and is not found in the original document's full-text representation. Our motivation is to assess the potential of citations as a source of new and relevant terms for document expansion. Intuitively, it would be interesting to find many new terms in citations, and for those terms to be related to the original terms. As a reference, we also built a baseline method where the sets of citations pointing to a target document were randomly assigned to a different document. Our aim with this baseline was to measure the amount of new and related terms that we would expect to find by chance from a random text snippet in the collection, and compare these numbers to the real citations to see if there is a clear signal.

Our approach to measuring the term relationships between document terms and citation terms consists of three steps: (i) identify all novel terms in the citation contexts (i.e. the terms not present in the original documents), (ii) for each term, obtain its synonyms and hypernym from the Metathesaurus, and (iii) search for these related words in the original representations; each match implies that the novel term in the citation has an ontological relationship to a term in the original document. This process allows us to identify the new citation terms that are synonyms and hypernyms of the terms in the original representation.

In order to define the terms to be used as unit of the analysis, we considered different approaches. We first explored the use of sliding windows to identify phrases present in the Metathesaurus. We tested windows up to three terms, and found that a large proportion of the matches were single tokens. We then applied the MetaMap[11] tool from the NLM to identify relevant phrases in the text; however we found that its phrase segmentation produced long strings containing UMLS concepts; and using those strings for look-up over the original documents would be problematic, and would produce an artificial increase in the amount of novel concepts found in citations. For instance, the phrase "heart size" can be identified by MetaMap in a citation, and looking up this phrase in the original document may not produce a match, even if "heart" and "size" are present, however we do not want to consider "heart size" as a novel concept.

A better way of using MetaMap would be to identify the substrings in the found phrases that belong to UMLS, but for simplicity in our work citation terms are defined as single

[11] http://metamap.nlm.nih.gov/

tokens, although the expansion terms (from the Metathesaurus) can be multi-words. The use of single words ensures that the terms identified as novel are new concepts not present in the original document, and not word ngrams

The results are shown in Table 2. We can see that most of the terms found in the citation contexts do not occur in the original, cited documents. Also an important percentage of those terms are synonyms or hypernyms of words in the original documents. In contrast, there are slightly more new terms in random citations, as expected, but less of these have related terms in the original document. We find less than half the amount of synonyms, and 37% less hypernyms. This suggests that citations can be a useful source of information.

We next look at the distribution of new terms and relationships within different logical sections in the scientific articles. Our goal is to measure if there are substantial differences according to the position of the citation in the text. For that, we segment each document into sections by relying on the headings. We identified eight main section names, and we map all the headings from all the papers into those eight categories using a set of manually generated rules. This is done by first listing all unique section headings using the HTML tags that delimit them; then examining the list manually and mapping each heading into one of the main headings. In cases where the mapping is not clear from the chosen words, we access the original paper, and map into the closest section heading after reading the content (e.g. "Data integration" into "Method"). Note that these cases were rare (less than 5% of the list).

After normalising the section headings, we analyse the distribution of citation terms in Table 3. The results show that section types -*Discussion*, *Introduction* and *Results*- contain the citation contexts with the highest proportion of terms that are semantically related to terms in the original document text. On the other hand, terms from *Conclusion* and *Future work* are scarce and less related. This information may be useful in the context of applications, and will be studied further in Section 7.

Our observations indicate that a large proportion of new and related terms (cf. Table 2) come from the top sections in which we find most of the citations (cf. Table 3). It is not surprising that most citation terms are found in sections, such as "*Discussion*", "*Introduc-*

tion", and "*Results*", as they are most commonly used by authors to compare their work and findings with other existing research. Authors might be expected to describe other related research using different words and terminologies in such sections; thus they are very likely to have new and related citation terms.

Likewise, sections like *Methods* and *Experiments* can be used to compare the current tools and methodologies with one another. These sections were found to have a large proportion of the new and related terms.

On the other hand, sections like "*Conclusion*" and "*Future work*" are less likely to be used to cite others. Rather, authors seem to use these sections to emphasise their findings and summarise their work (i.e. in "*Conclusion*"), and describe some work that they intend to complete (i.e. in "*Future work*").

4.2. Section weight and distance

We will focus now on the class distribution of terms as a way to measure their potential for text processing applications, such as clustering or text classification. Given a distribution of classes across documents, we expect the (class) discriminating power of a term to increase as class entropy lowers. We measure the discriminating power or "quality" of a term using *Shannon's entropy measurement* [46]. For all classes (i.e. MeSH tags), we compute the entropy of a term in our collection using the following equation:

$$H(t)=-\sum_{i=1}^{n} P(t_i)\log P(t_i) \tag{1}$$

where $P(t_i)$ is the probability that term t appears in class i, and n is the number of classes.

As explained earlier in Section 3, we rely on 20 MeSH terms to form the classes. To illustrate how class-entropy can be used to distinguish the most relevant terms of a given class, we show the top 20 terms ranked according to their entropy score (lowest first) in Table 4. In many cases, we can intuitively see why some terms have a strong relationship with certain classes. For example, the term “demography” appears in 37 documents belonging to class “Humans” out of 37 documents; whereas other classes have one or zero occurrences. Focusing on the major classes, for the class HUMAN the top terms in the list refer to information about studies (demography, ethnic, cohort, covariance, gender, multi-vari); the human body (forearm, supine); and human activities (smoke). While in the case of the class ANIMAL there are terms about habitat (forage, tank, freshwater, tidal); kinds of animals (trout, predator); and animal studies (thoracotomy, jugular, doppler, tunnel).

We will now use the class-entropy of terms to analyse two parameters: *Section* position, and *Distance* to the marker. To calculate the correlation coefficient between the class-entropy of terms and those parameters, we use the CORREL function, which calculates the Pearson Product-Moment Correlation Coefficient for two sets of values as follows:

$$CORREL(X,Y)=\frac{\sum(x-x^{-})(y-y^{-})}{\sqrt{\sum\left(x-x^{-}\right)^{2}\sum\left(y-y^{-}\right)^{2}}} \tag{2}$$

where x^{-} and y^{-} are the sample means of the x and y values, respectively.

Regarding the relationship between entropy and *Section* position, we define a section-score for each term, which measures the sections of the text it tends to occur in. The section weight is simply obtained by measuring the proportion of SYN and HYP terms found in the section (e.g. the section Discussion has a weight of 0.27, see Table 3 for further details). For every term, we calculate the average section weight as follows:

$$AW(t)=\frac{\sum_{i=1}^{n_t} W_{t,i}}{n_t} \tag{3}$$

where $AW(t)$ is the average weight of all sections in which term t appears, $W_{t,i}$ is the weight of section i containing t, (if t does not appear in any recognised section, $W_{t,i}$=0), and n_t is the number of occurrences of term t in the document.

Thus, for each term we calculate its class-entropy and average section weight. Next, we measure the correlation coefficient between the two parameters, obtaining a score of -0.46, which shows a strong negative correlation. For illustration, Figure 3 shows the relationship between a term's average section weight and its entropy. There seems to be a relationship between entropy and sections in which the terms occur, suggesting that terms with high average section score tend to have low entropy, and vice versa. This could indicate that sections with high scores (based on SYN and HYP density) tend to have the most valuable citation terms. This is a first indication that the section weight could be a relevant parameter for applying citation terms. For example, a term found in sections like "*Results*", "*Discussion*" and "*Introduction*" is likely to be more valuable than if it appears in a section like "*Conclusion*". This seems reasonable, as in general authors compare their work with related research within sections such as "*Discussion*" and"*Results*", whereas they tend to summarise their paper's contributions within the "*Conclusion*" section. We will explore this observation further in our text classification task (cf. Section 7) where we weight citation terms differently based on the sections in which they occur.

Finally, we explore the relationship between the entropy of citation terms with respect to their average *Distance* (in words) from their citation marker. For every term, we calculate the average distance (in words) as follows:

$$DW(t)=\frac{\sum_{i=1}^{n_t} D_{t,i}}{n_t} \tag{4}$$

where $DW(t)$ is the average distance of term t, $D_{t,i}$ is the number of terms between term t and citation marker i, and n_t is the number of occurrences of term t in the document.

Looking at Figure 4, in this case the correlation coefficient score is 0.14, which indicates that there is not clear linear relation between these values. This result may seem somewhat counter-intuitive.

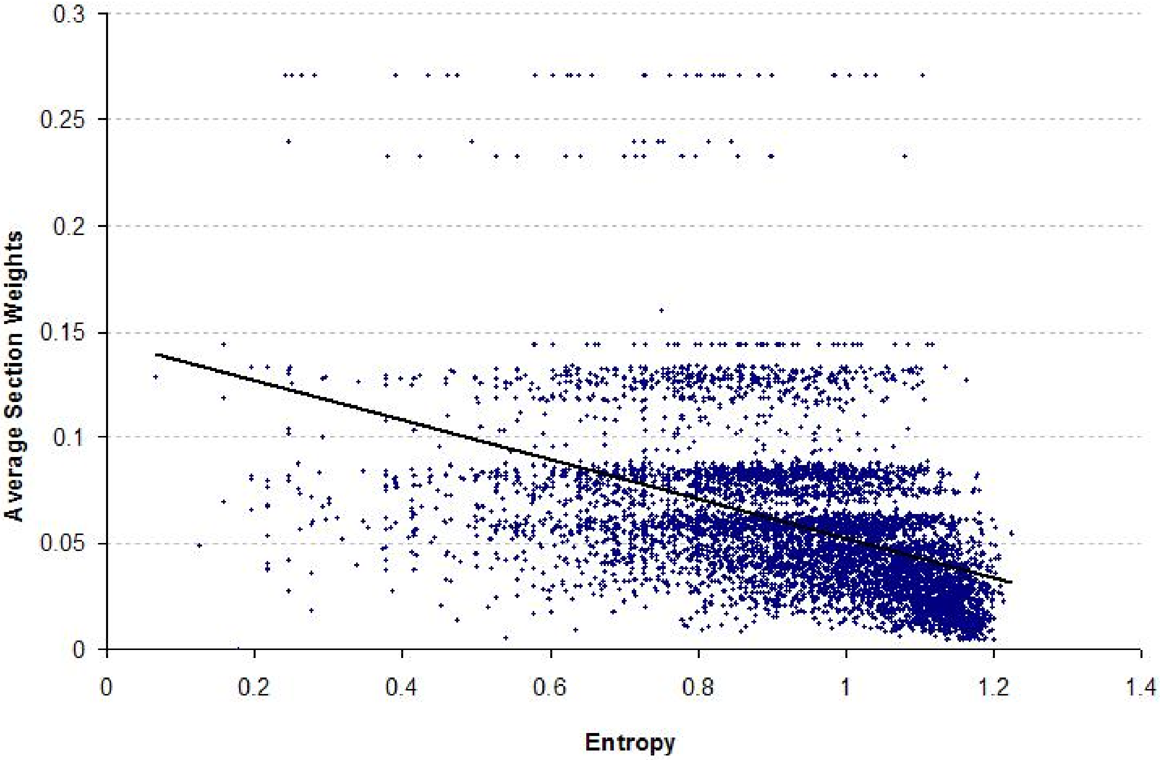


Figure 3: Graph showing the relationships between the Average weight of Sections and the Entropy of citation terms (with correlation coefficient of -0.46).

Generally speaking, in academic literature, there is no universal method which is used to cite others, so authors place citation markers in different positions even when the scope of citation is the same. For example, some authors start their citation context with citation markers, while others place citation markers at the end of citation contexts when they are finished discussing the related work. Some authors place the citation markers once they mention the work, then they continue to describe that work and other related findings. Alternatively, authors describe other work and compare it with theirs, and then they point to that work. Hence, the most "interesting" terms associated with the paper being cited are not necessarily closest to the citation marker. We will also test this parameter in our text classification experiments to confirm the usefulness of the distance information (cf. Section 7).

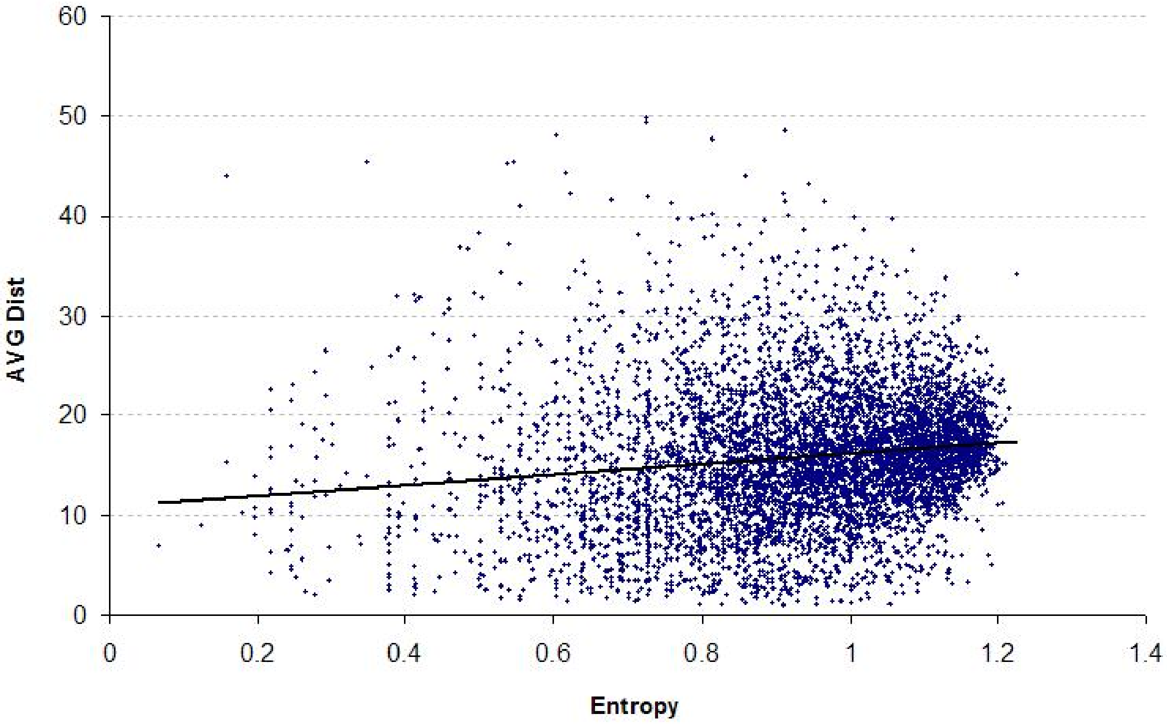


Figure 4: Graph showing the relationships between Distance to citation marker and the Entropy of citation terms (with correlation coefficient of 0.14).

5. Document representation for text classification

We now present an extrinsic evaluation of the "quality" of our citation terms using a text classification task, where the goal is to assign one or more semantic tags to each document, and compare the predictions to the manually-assigned tags. As described in Section 3, our target tags are MeSH terms, and our document collection is a subset of the TREC Genomics dataset. We explore different ways to model documents for this task, by relying on two resources: (i) the Metathesaurus, and (ii) citation contexts. In this section we first describe the different ways to enrich document representations, and then we explain weighting schemas for the terms.

5.1. Document enrichment

Document enrichment (also known as document expansion) is the process of adding related terms to the representation of the document. When measuring the similarity among documents, this technique can be used to overcome the problem of vocabulary mismatch, where a relevant document can be missed because a concept is referred to with a synonym. In IR for instance, document expansion techniques enrich documents off-line with related terms during indexing. This type of expansion can reduce the overhead of query expansion at query time. The drawback of this approach is that the ambiguity of query terms can introduce noise in the form of terms unrelated to the original sense of the query. In our work we attack this problem by combining two independent expansion sources: thesauri and citation contexts.

Thesaural Expansion: In thesaural or ontological based expansion, semantically related terms are obtained by looking up in the external resource. For instance, if the term *treatment* occurs in the original document, its synonym *intervention* can be added to the representation. We explore this option by extracting from the Metathesaurus all synonyms and hypernyms of the terms in the original document. For our basic approach we then incorporate these terms directly into the document representation, with the same frequency count as the original term.

In related work, Billerbeck and Zobel [47] proposed two new corpus-based methods for document expansion. In the first method, each document is treated as a query, and augmented by related terms. In the second method, each single term in the corpus is treated as a query, augmented by related terms, and used to rank documents accordingly. Overall, Billerbeck and Zobel's experiments showed that, compared with query expansion, document expansion methods achieved relatively poor improvements. That might be because the specific topic of the original document can be significantly skewed when less relevant related terms are added.

Citation Term Expansion: In this expansion strategy we gather the citation contexts that refer to the target document, and extract all terms occurring in those to expand the original representation. The motivation of this approach is twofold: (i) discover new terms that do not exist in the original representation, and (ii) boost the weight of the terms already found.

Combining Thesaural and Citation Term Expansion: In this expansion strategy we combine thesaural information with the terms from citation contexts. Our methodology is described in the following steps, and illustrated in Figure 5:

1. We first obtain the set of terms in the original representation of the document (D), and the terms that cite the document (C)
2. We obtain the set of novel terms (N) by selecting the citation terms that do not occur in D. ($N = C \setminus D$)
3. We expand N by obtaining all the synonyms and hypernyms of its terms in the UMLS database, and create a set of terms E. Note that these terms can be multiwords. ($E = \mathit{synonyms}(N) \cup \mathit{hypernyms}(N)$)
4. The expanded term set is reduced to those terms that do not occur in the original document D. ($E' = E \setminus D$)
5. Each term in the final expansion set *E'* is linked back to the term from C that originated it, and these pairs (c_i, e_i) of terms will be used for the final representation of the target documents.

We follow the above steps to build a set of pairs (c_i, e_i) for each document in the collection. These pairs will then be applied to build lookup dictionaries for expansion, which we call *citation dictionaries*. We implemented two different approaches, depending on the local or global use of the pair sets, which we describe below, and illustrate in Figure 6:

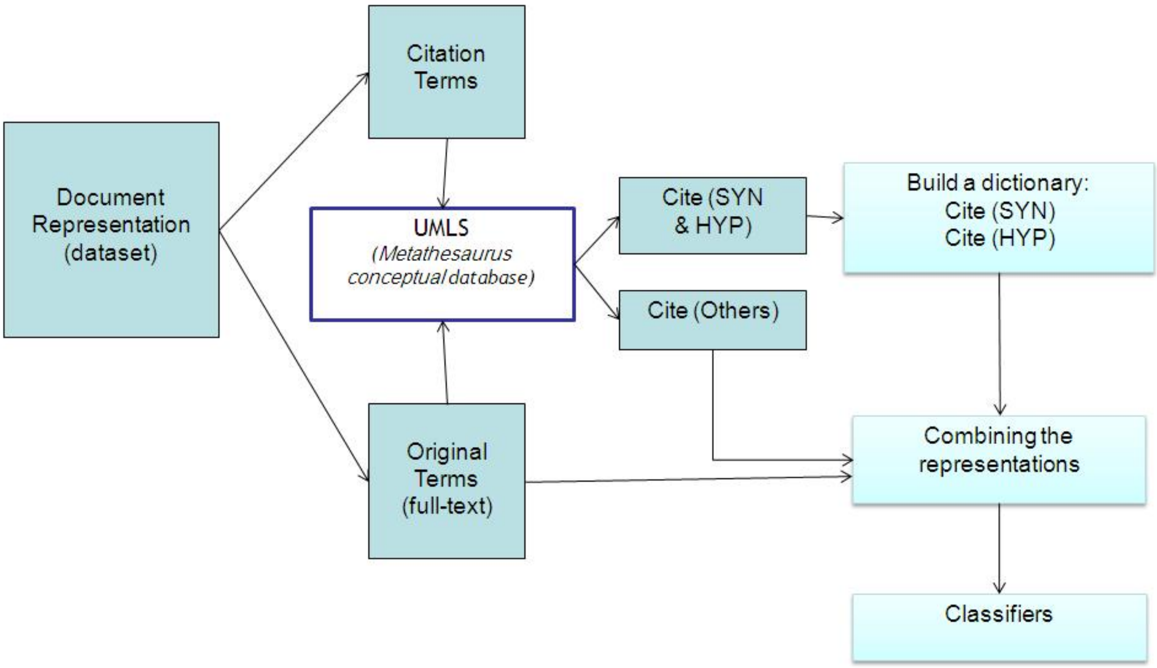
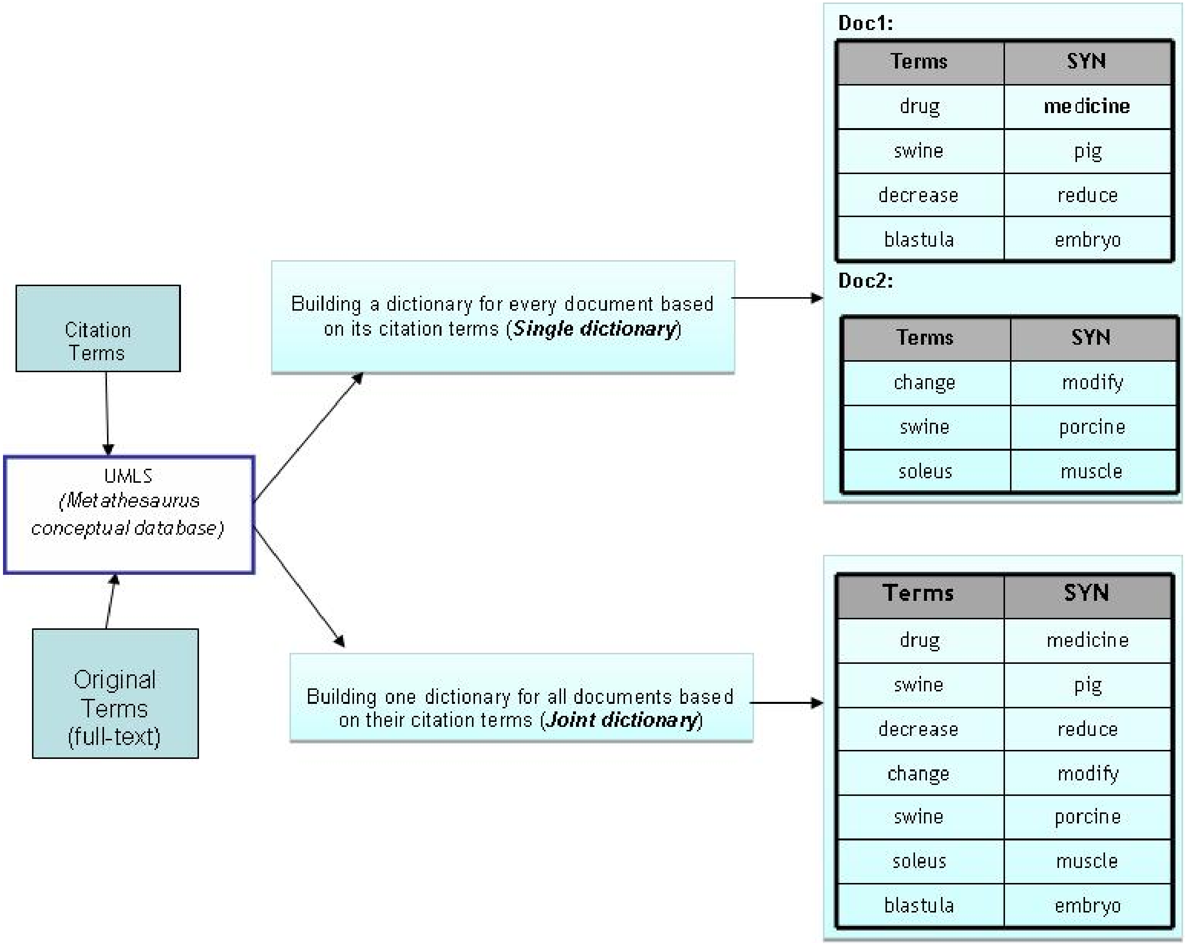

Citation Terms
Document Representation (dataset)
UMLS
(Metathesaurus conceptual database)
Cite (SYN & HYP)
Cite (Others)
Build a dictionary:
Cite (SYN)
Cite (HYP)
Original Terms (full-text)
Combining the representations
Classifiers

Figure 5: Graph showing our document expansion strategy using the Metathesaurus to filter out citation terms that hold no thesaural relationship with the original document terms.

- *Single dictionary*: We build a single lookup table (dictionary) for each document based only on the terms citing the target document. The synonyms and hypernyms identified in the process described above are used to populate the dictionary for the target document, and this dictionary is not shared. The advantage of building one related-term dictionary for each document is that expansion terms are more likely to be relevant to the document's topic given that all related terms are drawn solely from document's citation contexts. For instance, if we find the word "culture" in the document, thesauri expansion will use terms related to both "civilisation" and "laboratory culture"; however when we rely on this combined approach we require that the expansion terms occur both in citations and as related words. Therefore the terms related to "culture" will only be used for expansion if they are citing the target document, and if a paper receives citations regarding "laboratory culture" it is unlikely


Citation Terms
UMLS
(Metathesaurus conceptual database)
Original Terms (full-text)
Building a dictionary for every document based on its citation terms (Single dictionary)
Building one dictionary for all documents based on their citation terms (Joint dictionary)
Doc1:
Doc2:

Doc1:

Terms	SYN
drug	medicine
swine	pig
decrease	reduce
blastula	embryo

Doc2:

Terms	SYN
change	modify
swine	porcine
soleus	muscle

Terms	SYN
drug	medicine
swine	pig
decrease	reduce
change	modify
swine	porcine
soleus	muscle
blastula	embryo

Figure 6: Graph showing our *Single* and *Joint* document expansion strategies.

that it will also be cited regarding “civilisation”. The disadvantage of this strategy is that due to the MetaThesaurus filtering step, we can end up with a situation where documents have few or even zero related citation terms in their dictionaries, leading to minimal document expansion.

- *Joint dictionary*: We build one large lookup table based on all citation terms extracted for all documents in the collection. For each document, we collect citation terms and related words as in the previous case, but they are used to construct a single lookup table that it is shared among all target documents. This strategy nearly assures us that every document will be expanded with citation terms - and in some cases these citation terms will not have been extracted from their own citation contexts. In this way, the *Joint* dictionary can be viewed as a domain specific subset of the larger MetaThesaurus.

5.2. Term weighting schemas

For each document in our dataset we obtain two separate term-vectors generated a) from the original document and b) from the citation contexts. These vectors are merged into a combined representation. Many schemes have been proposed to derive the weights of each index term in the document representation vector. We apply the *tf-idf* feature weighting schema, where the term frequency is multiplied by the inverse document frequency. It is used to measure the weight of 'importance' of terms in a document. The *tf-idf* basically stands for the term frequency (*tf*) and the inverse document frequency (*idf*). The $tf_{i,j}$ (term frequency of t_i in document d_j) is defined as follows:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{5}$$

Where $n_{i,j}$ is the number of occurrences of the term t_i in document d_j, and the denominator is the sum of the number of appearances of all terms k in document d_j.

Thus, the idf_i (inverse document frequency of t_i in the corpus) is defined as follows:

$$idf_i = \log(|D|/|d_i|) \tag{6}$$

Where $|D|$ is total number of documents in the corpus, and $|d_i|$ is the number of documents in which term (t_i) appears. The final tf-idf *score is the product of the scores resulting from the previous two equations.*

Apart from feature weighting, we also experiment with feature selection (filtering). When the filter is activated, we remove the terms in our stopword lists[12], all terms that occur in more than 70% of documents, and all terms that occur in less than 1% of documents. When experimenting with abstracts alone, a lower threshold is used: terms that occur in less than 3 documents are removed.

In order to study different parameters, we modify the *tf* scheme by considering the section position and the distance of the term to its citation marker. Thus, for a given document, we follow these steps:

1. The basic *tf* scheme is applied to the original term vector.
2. The modified *tf* schemes are applied to its citation vector.
3. The two vectors are combined and the weights for the shared terms are calculated by adding the corresponding *tf* values for a term.

For the terms coming from *citations*, we propose modified tf scores based on two factors: (i) the section the term comes from, and (ii) the distance between the citation marker and the term. Thus, instead of a linear increase of the term frequency, we increase it non-linearly based on these factors.

Section based weighting scheme: We define *Section_tf* with the following formula:

$$Section\_tf(t) = \sum_{i=1}^{n_t}(1 + \alpha_{t,i}) \tag{7}$$

Where n_t is the number of occurrences of term t in the document, and $\alpha_{t,i}$ is the weight of the section i in which term t appears.

[12] Retrieved from http://www.cs.mu.oz.au/~jz/resources and from the Simple English Wikipedia (May 2008) http://simple.wikipedia.org/wiki/Wikipedia:Basic_English_alphabetical_wordlist

The section weight (α) is a density-based value taken from the statistics presented in Table 3, which showed the collection frequency of synonyms and hypernyms in particular sections of a document. For example, the section *Discussion* has about 27% of all synonyms and hypernyms, hence its weight is 0.27. The weight of the other sections are as follows: Introduction (0.24), Results (0.23), Methods (0.15), Experiments (0.16), Abstract (0.17), Conclusion (0.12), and Future work (0.15).

Distance based weighting scheme: Our second term weight modification strategy is calculated as the distance between the citation term and its citation marker, and is described by the following equation:

$$Distance\_tf(t) = \textstyle\sum_{i=1}^{n_t}(1 + \delta_{t,i}) \tag{8}$$

Where n_t is the number of occurrences of term t in the document, and $\delta_{t,i}$ is the weight calculated based on the distance between term t and citation marker i in a given document.

Th $\delta_{t,i}$ value is calculated as follows:

$$\delta_{t,i} = 1/dis_{t,i} \tag{9}$$

Where $dis_{t,i}$ the number of terms between term t and citation marker i, 1 if adjacent.

Thus, when the term is very close to the citation marker, it will get a higher weight than other citation terms that are further away.

Section and Distance based weighting scheme: Finally, we combine the two modified scores into a single value, with the following equation:

$$Section\&Distance\_tf(t) = \sum_{i=1}^{n_t}(1 + \alpha_{t,i} + \delta_{t,i}) \tag{10}$$

Where n_t is the number of occurrences of term t in the document, $\alpha_{t,i}$ is the weight of the section i in which term t appears, and $\delta_{t,i}$ is the weight calculated based on the distance between term t and citation marker i.

6. Text classification

We evaluate our methods extrinsically in the context of a supervised document classification task where documents are automatically assigned topic tags in the form of MeSH headings. As described in Section 3, we rely on a subset of the TREC Genomics dataset (3,475 documents) and the manually-assigned MeSH terms, focusing on the top-20. This is a multi-label classification problem, where each document will have one or more labels associated. Our goal is to develop and evaluate automatic classifiers to perform this task. Since we have access to both abstracts and full-text documents we compare the performance of our classification techniques on both collections.

We calculate the performance of the classification task based on Precision and Recall. Thus, for each class, Precision is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives). Recall is given by the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives). In order to combine these two scores into one, the F-score metric is used. F-score is the harmonic mean of Precision and Recall. Since our classification is a multi-class problem, and requires averaging all results from each class, we use the micro-averaging [41] method, which weights each class according to its number of instances. This is the usual approach when the errors from different classes have the same cost.

For comparison, we also include runs from two publicly available, state-of-the-art systems: MTI and MeSHUP, previously mentioned in Section 2.4. MTI is the NLM's currently deployed classification system, which uses the MetaMap concept parser for discovering MeSH headings. We use the system's default settings for MeSH classification, and its online interface[13]. MeSHUP, on the other hand, combines different ML and thesauri-based techniques into a hybrid classifier. For MeSHUP we use the open source implementation released by the authors. The input for these two tools is a fragment of text, and the output is a ranked list of MeSH terms. Since these systems work for all MeSH classes, we filter out tags not listed in our 20-class list. Finally, as an easier baseline, we apply the Majority Class approach, where each document is assigned the single most frequent class from training data.

For our own supervised classifier, we chose Support Vector Machines (SVM) for two main reasons [48]: i) the SVM performs well with large numbers of features, and ii) the SVM is especially helpful when there are few training samples in a multi-class classification task. In this paper we apply SVM using the implementation from the Weka toolkit [49], in which a document is represented by a vector of weighted terms. We rely on linear kernels and default parameters. In selected experiments we also apply the Naive Bayes classifier from the Weka toolkit, in order to see if there are relevant differences in performance. For all our experiments, we first build a separate binary classifier for each class, and the target document is assigned all classes tagged as positive.

To calculate the statistical significance of our results, we apply the Wilcoxon signed-rank test, which is a symmetric and non-parametric test. For two related samples, the Wilcoxon signed-rank test compares the differences between their measurements but does not need prior information about the form of the distribution of the measurements [50]. Hence, it is considered a useful alternative to the t-test when assumptions about the normal distribution of the data cannot be made.

[13] http://ii.nlm.nih.gov/mti.shtml

For our evaluation, we split randomly the dataset into two parts: two thirds for development, and the remainder as held-out test data. For the majority of the experiments we rely on the development dataset in ten-fold cross-validation. This development dataset is used to explore the effect of the different parameters, and the held-out data is kept untouched to avoid overfitting. In our final experiment we compare our main systems to the state of the art using the held-out test data.

In order to obtain the *Section* weights for the formula, we analyse both training and test instances, ignoring class labels. Our methodology is reminiscent of *Transductive* machine learning [51], or semi-supervised classification [52], both of which take advantage of unlabeled test data for building a model. In our case, *Section* information is a novel feature that reflects the location of citation term occurrences. In order to obtain more accurate estimations for this feature, we use the whole dataset to calculate the proportion of related terms (SYN and HYP) found in different sections. Calculation of this feature thus uses both training and test data, but does not use the class labels of either the training or test. So importantly, the class label information of the test instances is not being used when building the model.

7. Text classification results

For our first set of experiments we rely on the BOW representation, where only the terms in the original document are used, with no expansion. We present the results for the following configurations:

- Classifier used (SVM, Naive Bayes, MeSHUP or MTI)
- Source of terms (full text or abstract only)
- Feature selection (yes or no)

These results are given in Table 5. We can see that MTI performs poorly, while MeSHUP obtains much higher results and almost full recall. This result is consistent with the experiments reported for MTI and MeSHUP in [12]. MeSHUP performs well both with abstracts or full-text data, but SVM benefits from the full text. The best f-score is achieved by SVM when relying on full text, and feature selection; and this shows that our supervised approach is able to obtain state-of-the-art results over the development dataset. Naive Bayes obtains lower f-score than SVM overall, and we will rely on the latter as the baseline to explore the expansion techniques.

For our next experiment we evaluated the performance of different document enrichment approaches. For document representation, we use the BOW from the original document and expand it with the different strategies. Our baseline classifier is the best from the previous experiment: SVM trained over full text, with *tf-idf*, and feature selection. The expansion techniques rely on the following sources, which where described in Section 5.1:

- Citations: all the terms in the citations are added.
- MetaThesaurus: synonyms and hypernyms present in this knowledge base are used
- Combined dictionaries: citation terms are filtered according to the information in the Metathesaurus, generating individual and joint dictionaries.

We present the performance of the different expansions in Table 6. We can see that there are small improvements over the baseline, which are statistically significant according to the Wilcoxon signed-rank test. The best approaches overall are i) using all terms in citations, and ii) using the Joint dictionary based on synonyms. The expansions contribute to the precision of the classifier, and not the recall. This could happen because of our reliance on binary classifiers, which produce less false positives when they have expanded models.

For our next experiment, we combine citation terms with dictionary-based expansions. We present the results in Table 7. We can see that when using the joint-dictionary both the precision and recall of citation terms are improved, and we achieve the highest performance so far over this dataset.

In our next experiment, we analyse the effect of varying the *Distance* and *Section* position parameters on the performance of the citation terms as explained in Subsection 5.2. The results are presented in Table 8. Our *intrinsic* analysis (cf. Section 4) showed that there is no clear relationship between the quality of citation terms and their distance from the citation marker. Therefore, we expect no major improvement when this variable is considered in our experiments. In contrast, we find that *section quality* can influence the effectiveness of the citation terms. More specifically, when we boost the significance of terms that occur in important sections of the paper, a significant improvement can be achieved, reaching an f-score of 59.1%. This result is also consistent with the analysis performed in Section 4.

We then explore the effect of varying the window size boundary of the citation contexts. We tested the performance when using the full paragraph, and also different fixed windows (70, 50, 30, and 10 terms before and after the citation marker). These results are presented in Table 9. We can see that the window size is like the *Distance* parameter has no major effect, and the optimal window size appears to be around 50-terms.

To summarise our cross-validation results over the training data, we achieve our best performance using the SVM Citations+Joint method (with synonym based citation expansion, section, distance and window size of 50 parameters). This run achieves an f-score of 0.591; a statistically significant improvement over the baseline f-score of 0.575 which does not employ any citation context information in its feature representations.

In our final set of experiments, we apply our best SVM run configuration (SVM with citations and the joint dictionary with synonyms) to our test data. The results of these runs are presented in Table 10, the most important of which is that the expanded system outperforms two of the state-of-the-art classification systems, MeSHUP and MTI. It also significantly outperforms both SVM (baseline) and SVM (citations). These results confirm our original hypothesis that terms found in citation contexts can be used to enrich the document representations of the cited documents and improve text classification task performance; thus opening the way for better document representations for other applications. For this experiment we also show the performance per class in Figure 7, where we can see that most classes obtain improvements over the baseline, even though there are large performance differences depending on the target class.

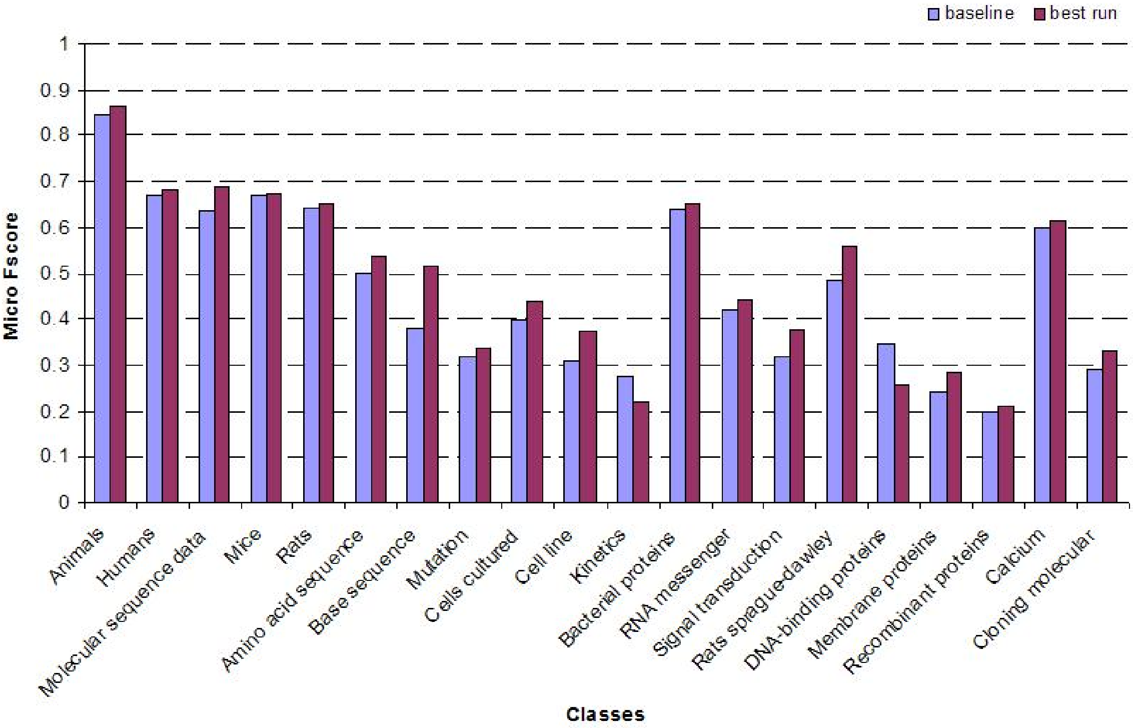


Figure 7: F-score over held-out data per class.

8. Findings

Our focus in this work was to empirically analyse the terms found in citation contexts, evaluate their quality (our intrinsic evaluation) and determine their effectiveness in a MeSH classification task (our extrinsic evaluation). Regarding the intrinsic evaluation, we observed that a high number of novel terms can be found, many of which are semantically related to terms in the original document. We also analysed two aspects of citation terms: (i) the section they are in, and (ii) the distance to the citation marker. We found that the section affects the quality of the citation terms, with some sections providing better terms than others (confirmed in both our intrinsic and extrinsic evaluation). On the other hand, the distance of citation terms to the marker (inside a fixed window) did not correlate with a term's quality (or performance).

Regarding the MeSH classification task, the following points can be drawn from the experiments presented in this paper:

1. Using citation terms as expansion terms is a promising strategy and can lead to improvements over baseline performance in a text classification task. In particular, weighting citation terms with respect to their *section* position in the citing document was found to have a very beneficial effect on our results. We were somewhat surprised to find that the distance from the *citation* marker was not as effective a method for weighting term importance. This result contradicts prior work in the area of Web IR and anchor text. This result was explored in both our *intrinsic* and *extrinsic* evaluation methods.
2. Using *synonyms* to expand documents tends to perform better than expansion with *hypernyms* (observed in 5 out of 6 experiments). This could be explained because synonyms naturally tend to be topically closer to the cited document than hypernyms do.
3. Our citation filtering, and section and distance weighting parameters appeared to stabilise the effects of varying the citation context window size around the marker. Specifically, these parameters medicate the possibility of adding unrelated citations terms to the document representation as the citation context window size increases by ensuring that only semantically related and highly weighted terms are considered.
4. Our most interesting finding is that our best citation expansion strategy involved using expansion terms derived from an automatically created domain specific dictionary - or the *Joint* dictionary. This dictionary was generated by first filtering out all citation terms that did not hold either a synonym or hypernym relationship with a term in the original cited document. This process was repeated for each document in the collection, and the set of remaining citation terms for each document was added to the *Joint* dictionary. Hence, this dictionary can be viewed as a specialised subset of the Metathesaurus, which captures concept relationships that are specific in to the Genomic domain. The idea of the *Joint* dictionary was motivated by our observation that after filtering, many documents were left without a corresponding set of citation

expansion terms - as was the case with the *Single* dictionary expansion strategy. With the *Joint* dictionary, this situation is corrected.

In summary, our results show that our expansion techniques can build a document model that significantly improves performance over state-of-the-art systems in a MeSH categorisation task. This is a strong indication that other text mining systems will also benefit from our document modeling method, resulting in improved performance of these systems in the biomedical domain.

9. Conclusions

In this paper, we investigate the different factors influencing the use of citation terms as a means of enriching the representation of a document with additional informative synonyms and related terms. First, we conducted an *intrinsic* evaluation which explored the types of relationships that exist between original document terms and terms found in the citation contexts that refer to them. More specifically, we analysed the terms from citation contexts in our collection and found that they are a rich source of topically related terms, i.e. synonyms and hypernyms. Interestingly, these terms are in general not found in the original full-text versions of the scientific articles that we examined.

Next we employed an *extrinsic* evaluation method which relies on an automatic classification of MeSH terms for MedLine documents. In these experiments, we explored the effect of document enrichment using citation terms and ontological terms (taken from the UMLS Metathesaurus). However, only small increments in performance were observed when the latter were considered. Our final experimental run combined both citation terms and Metathesaurus using only citation terms. Classification experiments with this enriched document representation achieve a statistically significant improvement over both the baseline and two state-of-the-art MeSH classification systems: MTI and MeSHUP.

We also explore the different factors affecting citation term effectiveness, including section position in the text, and distance from the citation marker, as well as the optimal window size of the citation context boundary. The section-based weighting scheme showed some improvement gains, indicating that consideration of document structure may be an interesting avenue for future work. However, the distance metric did not provide any noticeable improvements, which does seem to contradict related work in Web information retrieval where anchor text terms closer to the hypertext link have been shown to be more topically relevant to the linked page.

In future work, we plan to explore alternative sources of related terms including: n-gram term co-occurrence analysis, and the other hierarchical thesaural relationship types. We also plan to explore different Section (α) and Distance (δ) weighting techniques. The analysis of features coming from different sources is also in our agenda. Finally, we would like to study the automatic classification of citation boundaries for a more accurate selection of terms.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] P. Nakov, A. Schwartz, M. Hearst, Citances: Citation sentences for semantic analysis of bioscience text, in: Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics, Sheffield, UK, 2004, pp. 81–88.

[2] P. Wanjantuk, J. Keane, Finding related documents via communities in the citation graph, in: Proceedings of the International Symposium on Communications and Information Technologies, (ISCIT), Sapporo, Japan, 2004, pp. 445–450.

[3] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, D. Radev, Blind men and elephants: What do citation summaries tell us about a research article?, J. Am. Soc. Inf. Sci. Technol. 59 (1) (2008) 51–62. doi:http://dx.doi.org/10.1002/asi.v59:1.

[4] A. Ritchie, S. Robertson, S. Teufel, Comparing citation contexts for information retrieval, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), ACM, Napa Valley, California, USA, 2008, pp. 213–222. doi:http://doi.acm.org/10.1145/1458082.1458113.

[5] A. Ritchie, S. Teufel, S. Robertson, Using terms from citations for ir: some first results, in: Proceedings of the IR research, 30th European conference on Advances in Information Retrieval (ECIR), Springer-Verlag, Glasgow, UK, 2008, pp. 211–221.

[6] L. Hunter, K. Cohen, Biomedical language processing: What's beyond pubmed?, Molecular cell 21 (5) (2006) 589–594.

[7] R. Rak, L. Kurgan, M. Reformat, Multilabel associative classification categorization of MEDLINE articles into MeSH keywords, IEEE Eng Med Biol Mag 26 (2) (2007) 47–55.

[8] S. Scott, S. Matwin, Feature engineering for text classification, in: Proceedings of 16th International Conference on Machine Learning (ICML), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 379–388.

[9] A. McCallum, R. Rosenfeld, T. Mitchell, A. Ng, Improving text classification by shrinkage in a hierarchy of classes, in: Proceedings of 15th International Conference on Machine Learning (ICML), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 359–367.

[10] W. Lam, C. Ho, Using a generalized instance set for automatic text categorization, in: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), ACM, Melbourne, Australia, 1998, pp. 81–89. doi:http://doi.acm.org/10.1145/290941.290961.

[11] G. Kim, A. Aronson, J. Mork, B. Cohen, C. Lehmann, Application of a medical text indexer to an online dermatology atlas, Stud Health Technol Inform 107 (1) (2004) 287–291.

[12] D. Trieschnigg, P. Pezik, V. Lee, F. Jong, W. Kraaij, D. Rebholz-Schuhman, MeSH Up: effective MeSH text classification for improved document retrieval, BIOINFORMATICS 25 (11) (2009) 1412–1418.

[13] E. Garfield, Information, power, and the science citation index, Essays of an Information Scientist 1 (1972) 266–267, institute for Scientific Information.

[14] R. Mercer, C. D. Marco, A design methodology for a biomedical literature indexing tool using the rhetoric of science, in: Proceedings of the BioLink workshop in conjunction with (HLT/NAACL), Linking Biological Literature, Ontologies and Databases, Boston, MA, Association for Computational Linguistics, 2004, pp. 77–84.

[15] H. Nanba, N. Kando, M. Okumura, Towards multi paper summarization using reference information, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 926–931.

[16] S. Teufel, M. Moens, Summarizing scientific articles: Experiments with relevance and rhetorical status, Computational Linguistics 28 (4) (2002) 409–445.

[17] S. Teufel, A. Siddharthan, C. Batchelor, Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3, EMNLP '09, Association for

Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 1493–1502.
URL `http://portal.acm.org.ezp.lib.unimelb.edu.au/citation.cfm?id=1699648.1699696`

[18] H. Nanba, T. Abekawa, M. Okumura, S. Saito, Bilingual PRESRI integration of multiple research paper databases, in: Proceedings of the Conference on Large-Scale Semantic Access to Content (RIAO), Avignon, France, 2004, pp. 195–211.

[19] H. Nanba, N. Kando, M. Okumura, Classification of research papers using citation links and citation types: Towards automatic review article generation, in: Proceedings of The 11th SIG Classification Research Workshop, Classification for User Support and Learning, 2000, pp. 117–134.

[20] H. Nanba, M. Okumura, Automatic detection of survey articles, in: A. Rauber, S. Christodoulakis, A. M. Tjoa (Eds.), Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL 2005, Vienna, Austria, September 18-23, 2005, Proceedings, Vol. 3652 of Lecture Notes in Computer Science, Springer, 2005, pp. 391–401.

[21] S. Teufel, Argumentative Zoning: Information extraction from scientific text, Ph.D. thesis, Edinburgh, UK (1999).

[22] A. Ritchie, S. Teufel, S. Robertson, How to find better index terms through citations, in: Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval? (CLIIR), Association for Computational Linguistics, Sydney, Australia, 2006, pp. 25–32.

[23] S. Mohammad, B. Dorr, M. Egan, A. Hassan, P. Muthukrishan, V. Qazvinian, D. Radev, D. Zajic, Using citations to generate surveys of scientific paradigms, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 584–592.

[24] S. Wan, C. Paris, R. Dale, Whetting the appetite of scientists: producing summaries tailored to the citation context, in: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), ACM, Austin, TX, USA, 2009, pp. 59–68. doi:http://doi.acm.org/10.1145/1555400.1555410.

[25] S. Wan, C. Paris, R. Dale, Invited paper: Supporting browsing-specific information needs: Introducing the citation-sensitive in-browser summariser, Web Semant. 8 (2-3) (2010) 196–202. doi:http://dx.doi.org/10.1016/j.websem.2010.03.002.

[26] S. Wan, C. Paris, M. Muthukrishna, R. Dale, Designing a citation-sensitive research tool: an initial study of browsing-specific information needs, in: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL), Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 45–53.

[27] S. Bradshaw, Reference directed indexing: Redeeming relevance for subject search in citation indexes, in: Proceedings of the 7th European Conference on Research and Advanced Technology for Digital

Libraries (ECDL), Springer Berlin / Heidelberg, Trondheim, Norway, 2003, pp. 499–510.

[28] S. Bradshaw, Reference directed indexing: Indexing scientific literature in the context of its use.

[29] B. Aljaber, N. Stokes, J. Bailey, J. Pei, Document clustering of scientific texts using citation contexts, Information Retrieval 13 (2) (2010) 101–131. doi:http://dx.doi.org/10.1007/s10791-009-9108-x.

[30] S. Bradshaw, Document indexing vocabularies: Reference vs content, Northwestern University (Technical Report, NWU-CS-01-7).

[31] S. Bradshaw, K. Hammond, Automatically indexing documents: content vs. reference, in: Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI), ACM, San Francisco, California, USA, 2002, pp. 180–181. doi:http://doi.acm.org/10.1145/502716.502746.

[32] K. Sugiyama, T. Kumar, M. Kan, R. Tripathi, Identifying citing sentences in research papers using supervised learning, in: Proceedings of the International Conference on Information Retrieval and Knowledge Management (CAMP), Shah Alam, Malaysia, 2010, pp. 67–72.

[33] Q. He, J. Pei, D. Kifer, P. Mitra, L. Giles, Context-aware citation recommendation, in: Proceedings of the 19th international conference on World Wide Web (WWW), ACM, Raleigh, North Carolina, USA, 2010, pp. 421–430. doi:http://doi.acm.org/10.1145/1772690.1772734.

[34] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, G. Flake, Using web structure for classifying and describing web pages, in: Proceedings of the 11th international conference on World Wide Web (WWW), ACM, Honolulu, Hawaii, USA, 2002, pp. 562–569. doi:http://doi.acm.org/10.1145/511446.511520.

[35] S. Lawrence, C. Giles, K. Bollacker, Digital libraries and autonomous citation indexing, Computer 32 (6) (1999) 67–71. doi:http://dx.doi.org/10.1109/2.769447.

[36] C. Giles, K. Bollacker, S. Lawrence, Citeseer: an automatic citation indexing system, in: Proceedings of the 3rd ACM Conference on Digital Libraries (DL), ACM, Pittsburgh, Pennsylvania, United States, 1998, pp. 89–98. doi:http://doi.acm.org/10.1145/276675.276685.

[37] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, in: Proceedings of the 7th International Conference on World Wide Web (WWW), Brisbane, Australia, 1998, pp. 107–117.

[38] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632. doi:http://doi.acm.org/10.1145/324133.324140.

[39] M. Koolen, J. Kamps, The importance of anchor text for ad hoc search revisited, in: Proceedings of the 33rd International ACM Conference on Research and Development in Information Retrieval (SIGIR), ACM, Geneva, Switzerland, 2010, pp. 122–129. doi:http://doi.acm.org/10.1145/1835449.1835472.

[40] P. Dobrokhotov, B. Goutte, C. Veuthey, E. Gaussier, Combining NLP and probabilistic categorisation of document and term selection for Swiss-Prot medical annotation, Bioinformatics 19 (1) (2003) 91–94.

[41] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002)

1–47. doi:http://doi.acm.org/10.1145/505282.505283.

[42] F. Liu, T. Jenssen, V. Nygaard, Figsearch: A figure legend indexing and classification system, Bioinformatics 20 (16) (2004) 2880–2882.

[43] A. Cohen, W. Hersh, A survey of current work in biomedical text mining, BRIEFINGS IN BIOINFORMATICS 6 (1) (2005) 57–71.

[44] S. Sohn, W. Kim, D. Comeau, W. Wilbur, Optimal training sets for bayesian prediction of MeSH assignment, J Am Med Inform Assoc 15 (4) (2008) 546–553.

[45] N. Shah, M. Musen, UMLS-Query: A perl module for querying the UMLS, in: AMIA Annual Symposium, 2008, pp. 652–656.

[46] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.

[47] B. Billerbeck, J. Zobel, Document expansion versus query expansion for ad-hoc retrieval, in: Proceedings of the 10th Austrasian Document Computing Symposium (ADCS), Sydney, Australia, 2005, pp. 34–41.

[48] S. Mukherjee, Classifying microarray data using support vector machines, in: W. D. D.P. Berrar, M. Granzow (Eds.), A Practical Approach to Microarray Data Analysis, Kluwer Academic Publishers, Boston, MA, 2003, Ch. 9, pp. 166–185.

[49] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[50] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (Dec., 1945) 80–83.

[51] V. Vapnik, Statistical learning theory, in: A. Gammerman (Ed.), Computational Learning and Probabilistic Reasoning, Wiley, Qld, Australia, 1996, Ch. 24.

[52] V. Vapnik, Transductive inference and semi-supervised learning, in: O. Chapelle, B. Scholkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT press, Cambridge, MA, USA, Ch. 24, p. 454.

Rank	MeSH term	# Total Freq˙	Development	Held-out
1	Animals	2086	1434	652
2	Humans	1206	809	397
3	Molecular sequence data	690	464	226
4	Mice	606	404	202
5	Rats	593	395	198
6	Amino acid sequence	508	346	162
7	Base sequence	389	264	125
8	Mutation	361	252	109
9	Cells cultured	344	230	114
10	Cell line	334	235	99
11	Kinetics	299	200	99
12	Bacterial proteins	296	197	99
13	RNA messenger	292	197	95
14	Signal transduction	285	194	91
15	Rats sprague-dawley	245	163	82
16	DNA-binding proteins	234	156	78
17	Membrane proteins	223	150	73
18	Recombinant proteins	217	145	72
19	Calcium	211	141	70
20	Cloning molecular	197	131	66

Table 1: 20 most frequent MeSH terms in our collection, and their document-frequencies in our corpus.

	Random-Citation		Real-Citation	
Terms	#	%	#	%
All	30,240	100	30,240	100
New	28,047	92.7	26,512	87.7
Synonyms	2,057	7.3	4,149	15.6
Hypernyms	1,203	4.2	1,904	7.2

Table 2: Semantic analysis of words found in citation contexts. The percentages of synonyms and hypernyms are calculated over the set of new terms only.

Sections	All terms	% New	% SYN	% HYP
Discussion	**15,717**	83.4	**18.6**	**8.5**
Introduction	14,063	78.8	16.7	7.2
Results	12,712	82.0	16.0	7.3
Methods	10,062	**84.5**	9.8	4.7
Experiments	3,976	**84.5**	10.2	5.7
Abstract	2,090	70.7	11.7	5.0
Conclusion	638	66.3	7.8	3.8
Future work	223	44.4	13.1	2.0

Table 3: Semantic analysis of words found in citation contexts per section. For each section, we provide the number of unique terms, the percentage of new terms, and the percentages of semantically-related terms (synonyms and hypernyms) in the new terms. The highest numbers per column are given in bold.

Order	Term	Entropy	# Classes	# Docs	Top ordered classes (# occurrences)
1	demography	0.284	7	37	Humans(37); Molecular(1); RNA, Messenger(1); Animals(1) Rats(1)
2	forage	0.589	10	34	Animals(31); Mice(3); Molecular Sequence Data(3); Base Sequence(2); Bacterial Proteins(2)
3	Doppler	0.610	8	37	Animals(26); Humans(11); Mice(5); Rats(2); Kinetics(2)
4	smoke	0.665	13	64	Humans(57); Animals(10); RNA, Messenger(6); Cells, Cultured(5); Kinetics(3);
5	ethnic	0.671	11	40	Humans (38); Mutation(5); Animals(5); Mice(4); Base Sequence(3);
6	trout	0.687	10	35	Animals(34); Molecular Sequence Data(5); Amino Acid Sequence(5); Kinetics(5); Base Sequence(4);
7	forearm	0.716	13	37	Humans(31); Animals(9); Rats(4); Kinetics(3); Rats, Sprague-Dawley(2)
8	predator	0.727	13	50	Animals(47); Humans(8); Molecular Sequence Data(8); Amino Acid Sequence(6); Mice(5); Base Sequence(3);
9	Thoracotomy	0.730	10	43	Animals(42); Mice(13); Rats(10); RNA, Messenger(5); Calcium(4);
10	supine	0.730	11	55	Humans(34); Animals(22); Mice(9); Kinetics(4); Rats(4);
11	cohort	0.741	15	94	Humans(74); Animals(26); Mice(10); Mutation(6); Rats(5);
12	covariance	0.746	11	49	Humans(28); Animals(19); Mice(9); Kinetics(4); Amino Acid Sequence(3);
13	gender	0.755	15	86	Humans (52); Animals(44); Mice(25); Cells, Cultured(6); Rats(4); RNA, Messenger(3);
14	tidal	0.760	12	40	Animals(37); Mice(13); Humans (8); Rats(7); Rats, Sprague-Dawley(4);
15	jugular	0.768	13	78	Animals(76); Rats(38); Rats, Sprague-Dawley(26); Mice(23); Humans(11); RNA, Messenger(7);
16	Multi-Vari	0.773	14	49	Humans(33); Animals(15); Mice(4); Bacterial Proteins(3); RNA, Messenger(3);
17	tank	0.778	15	45	Animals(37); Humans(10); Molecular Sequence Data(3); Amino Acid Sequence(3); Rats(3)
18	freshwater	0.779	12	34	Animals(31); Molecular Sequence Data(7); Amino Acid Sequence(6); Mutation(4); Membrane Proteins(4);
19	tunnel	0.785	14	36	Animals (32); Mice(12); Rats(7); Humans(4); Molecular Sequence Data(2)
20	Hyperinsulinemia	0.794	13	61	Animals(40); Humans(30); Rats(23); Rats, Sprague-Dawley(12); Kinetics(8); Mice(4);

Table 4: The top-20 terms ranked according to entropy (lowest first) and frequency. The terms are taken from the original document's full-text representation.

System	Ft. Sel.	Full-text			Abstract		
		P	R	F	P	R	F
Majority Class	-	0.142	0.220	0.173	0.142	0.220	0.173
MeSHUP	-	0.399	**0.978**	0.567	0.403	**0.966**	**0.569**
MTI	-	0.515	0.319	0.394	0.526	0.257	0.346
Naive Bayes	No	0.526	0.610	0.565	0.457	0.628	0.529
Naive Bayes	Yes	0.537	0.582	0.559	0.455	0.626	0.527
SVM	No	0.597	0.518	0.555	**0.567**	0.505	0.534
SVM	Yes	**0.610**	0.543	**0.575**	0.560	0.504	0.531

Table 5: Performance of classifiers when relying on BOW for feature representation (Ft. Sel.: Feature Selection, Ft. Weight: Feature Weighting). The best result per column is highlighted in bold.

System	P	R	F
Baseline	0.610	**0.543**	0.575
Citations	**0.637‡**	0.535	**0.582‡**
MetaThesaurus (syn)	0.629‡	0.533	0.577
MetaThesaurus (hyp)	0.617	0.538	0.575
MetaThesaurus (syn&hyp)	0.627‡	0.529	0.574
Single-dic (syn)	0.615	0.538	0.574
Single-dic (hyp)	0.616	0.542	0.577
Single-dic (syn&hyp)	0.622‡	0.540	0.578
Joint-dic (syn)	0.626‡	0.541	**0.581‡**
Joint-dic (hyp)	0.623‡	0.536	0.577
Joint-dic (syn&hyp)	0.627‡	0.531	0.575

Table 6: Performance of SVM using different document enrichment strategies. All statistical significance improvements over the SVM (baseline) are indicated by ‡(<=0.05).

System	P	R	F
Citations+MetaThesaurus(syn)	0.637‡	0.527	0.577
Citations+MetaThesaurus(hyp)	0.629‡	0.528	0.574
Citations+MetaThesaurus(syn&hyp)	0.637‡	0.526	0.576
Citations+Single-dic(syn)	0.628‡	0.531	0.576
Citations+Single-dic(hyp)	0.625‡	0.532	0.575
Citations+Single-dic(syn&hyp)	0.630‡	0.530	0.576
Citations+Joint-dic(syn)	**0.640‡**	**0.539**	**0.585‡**
Citations+Joint-dic(hyp)	0.630‡	0.531	0.576
Citations+Joint-dic(syn&hyp)	0.634‡	0.527	0.575

Table 7: Performance of SVM after combining citations and other document expansions. Statistical significance over original result indicated by ‡(<=0.05).

System	Section (α)	Distance (δ)	P	R	F
Citations+Joint-dic(syn)	N	N	0.640‡	0.539	0.585‡
Citations+Joint-dic(syn)	N	Y	0.643‡	0.539	0.587‡
Citations+Joint-dic(syn)	Y	N	**0.654‡***	0.539	**0.591‡**
Citations+Joint-dic(syn)	Y	Y	**0.654‡***	**0.540**	**0.591‡**

Table 8: Performance of SVM with Citations+Joint-dic(syn) according to the distance and section parameters. Statistical significance over original result indicated by ‡; and over Citations+Joint-dic(syn) result indicated by * (<=0.05).

System	Window Size	P	R	F
Citations+Joint-dic(syn)	Full Paragraph	0.653‡	**0.540**	**0.591**‡
Citations+Joint-dic(syn)	70	0.653‡	0.539	0.590‡
Citations+Joint-dic(syn)	50	**0.654**‡	**0.540**	**0.591**‡
Citations+Joint-dic(syn)	30	0.652‡	0.539	0.590‡
Citations+Joint-dic(syn)	10	0.645‡	0.537	0.586‡

Table 9: Performance of our classifiers after combining citation and Metathesaurus (based on citation terms only) expansions. Statistical significance over original result indicated by ‡(<=0.05).

System	P	R	F
MeSHUP	0.396	**0.976**	0.563
MTI	0.559	0.334	0.417
SVM (baseline)	0.606	0.548	0.576
SVM (citations)	0.635‡	0.538	0.582
SVM (Citation+Joint-dic(syn))	**0.665**‡*	0.553	**0.604**‡*

Table 10: Performance of optimised text classification runs on test data. Statistical significance over MeSHUP run indicated by ‡; and over SVM (citations) result indicated by * (<=0.05).