ITEXT-BIO: Intelligent Term EXTraction for BIOmedical Analysis

Rodrique Kafando, Rémy Decoupes, Sarah Valentin, Lucile Sautot, Maguelonne Teisseire, Mathieu Roche

Noname manuscript No.
(will be inserted by the editor)

ITEXT-BIO: Intelligent Term EXTraction for BIOmedical Analysis

Rodrique Kafando · Rémy Decoupes · Sarah Valentin · Lucile Sautot · Maguelonne Teisseire · Mathieu Roche

Received: date / Accepted: date

Abstract Here, we introduce ITEXT-BIO, an intelligent process for biomedical domain terminology extraction from textual documents and subsequent analysis. The proposed methodology consists of two complementary approaches, including free and driven term extraction. The first is based on term extraction with statistical measures, while the second considers morphosyntactic variation rules to extract term variants from the corpus.

The combination of two term extraction and analysis strategies is the keystone of ITEXT-BIO. These include combined intra-corpus strategies that enable term extraction and analysis either from a single corpus (intra), or from corpora (inter). We assessed the two approaches, the corpus or corpora to be analysed and the type of statistical measures used.

Our experimental findings revealed that the proposed methodology could be used: 1) to efficiently extract representative, discriminant and new terms from a given corpus or corpora, and 2) to provide quantitative and qualitative analyses on these terms regarding the study domain.

Keywords Biomedical terminology · Terminology extraction · Intelligent analysis

Rodrique Kafando · Rémy Decoupes · Maguelonne Teisseire
INRAE, Montpellier, France
TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
E-mail: firstname.lastname@inrae.fr

Sarah Valentin
CIRAD, F-34398 Montpellier, France
ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France
TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
E-mail: firstname.lastname@cirad.fr

Lucile Sautot
AgroParisTech, Montpellier - France
TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
E-mail: lucile.sautot@agroparistech.fr

Mathieu Roche
CIRAD, F-34398 Montpellier, France
TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
E-mail: firstname.lastname@cirad.fr

1 Introduction

The usefulness of terminology extraction from corpora is clearly acknowledged as it has generated a great deal of research and discussion. This well-established process is used in natural language processing and has led to the development of several tailored tools such as TBXTools (Oliver and Vàzquez, 2015), TermSuite (Cram and Daille, 2016), BioTex (Lossio-Ventura et al., 2014b), etc.

Based on (Lossio-Ventura et al., 2014b), our proposal deals with domain-based terminology extraction from heterogeneous corpora, and how to efficiently generate a quantitative and qualitative analysis. To this end, we propose a generic methodology hinged on a combination of extraction and analysis strategies. Term extraction strategies are based on combinations of linguistic, statistical measures, and corpus segmentation approaches, while analysis strategies are based on combinations of extracted terms.

Based on the combined strategies, ITEXT-BIO aims to extract: 1) representative terms, 2) discriminant or relevant terms, and 3) new relevant terms from a corpus or corpora. These strategies are specifically useful for dedicated tasks, such as corpus analysis, specific

domain monitoring (e.g. epidemiology) or scientific research monitoring.

This paper is organized as follows. In section 2, we briefly present the state-of-the-art related to terminology extraction. Section 3 details the dataset dedicated to scientific papers. Sections 4 and 5 respectively provide an overview of our proposal and the experiments. In section 6, we illustrate the genericity of the proposal by presenting a case study of an implementation of the combined strategies for epidemiological intelligence analysis. We conclude in section 7 by presenting some perspectives for future studies.

2 Related work

Domain terminology extraction is a major focus of interest and discussion in natural language processing (NLP) research. It has prompted several proposals of methodologies (Kageura and Umino, 1996; Pais and Ion, 2020; Pazienza et al., 2005; Rigouts Terryn et al., 2020) geared towards effective extraction of terms within a given corpus. Also known as automatic term extraction (ATE), this task is considered in various NLP applications, such as in information retrieval (Bracewell et al., 2005; Azarafza et al., 2020; Shah et al., 2019; Duari and Bhatnagar, 2020), topic modeling (Habibi and Popescu-Belis, 2015; Wang et al., 2020b), domain-based monitoring (Maynard et al., 2005; Joung and Kim, 2017; Arsevska et al., 2018), keyword extraction (Campos et al., 2020) and summarization (Azarafza et al., 2020), ontology acquisition, thesaurus construction, etc.

According to (Lossio-Ventura et al., 2014c), term extraction techniques can be categorized under four approaches: linguistic, statistical, machine learning and hybrid.

Overall, linguistic approaches take morphosyntactic part-of-speach (POS) rules into account to describe terms with common structures (Brill, 1992). Statistical approaches use statistical measures such as term frequency (Ramos et al., 2003; Whissell and Clarke, 2011), or term co-occurrence between words and phrases like Chi-square (Matsuo and Ishizuka, 2004). Machine learning approaches use statistical measures and are mainly jointly focused on term extraction (Conrado et al., 2013; Foo, 2009; Campos et al., 2020), classification (Wang et al., 2016) and summarization (Azarafza et al., 2020). They combine linguistic and statistic approaches to extract terms from textual data in order to build machine learning models. In (Campos et al., 2020), the authors highlighted that most of these tasks are tackled with unsupervised learning algorithms. Hybrid approaches include, for instance, C_Value (Pazienza et al., 2005), C/NC_Value (Frantzi et al., 2000) methods, which combine statistical measures and linguistic based rules to extract multi-word and nested terms. In (Campillos Llanos et al., 2013; Neifar et al., 2016), the authors combine rule-based methods and dictionaries to extract terms from Spanish biomedical texts and specialised Arabic texts respectively.

Studies such as (Ji et al., 2007; Lossio-Ventura et al., 2014a) related to these latter approaches have revealed the effectiveness and high performance of hybrid term extraction approaches.

The proposed methodologies apply to several domains. In (Lossio-Ventura et al., 2014b), the authors proposed BioTex, a linguistic and statistical measure-based tool to extract terms related to the biomedical domain. The same approach was used in (Arsevska et al., 2018) to detect terms or signals for infectious disease monitoring on the web. In (Meystre et al., 2008), a hybrid methodology was proposed to extract terminology for electronic heath records. This hybrid approach was also adapted by (Yao et al., 2017) to extract concepts related to Chinese culture.

The overall related studies have focused on techniques and methods for term extraction mainly from corpora. Based on existing methodologies, we oriented our study to develop an efficient approach for term extraction from heterogeneous corpora, along with a set of combined strategies to analyze these terms in the biomedical domain. Our methodology combines and tailors linguistic and statistic criteria associated with structural information in texts in order to highlight relevant terms therein. The presented strategies also aim to overcome the time-consuming issues related to machine learning methods which require manually annotated or partially annotated data.

3 Dataset description

Our study focused on the COVID-19 Open Research Dataset[1] (Wang et al., 2020a) which contains scientific papers on COVID-19 and related historical coronavirus research. Throughout this study, we refer to the dataset as COVID19-MOOD-data.

The COVID19-MOOD-data dataset is divided into two main corpora, respectively named Papers1 and Papers2. Papers1 contains the *commercial use subset (includes PubMed Central content)*, while Papers2 contains the *commercial use subset (includes PubMed Central content), the non-commercial use subset (includes PubMed Central content) and the custom license subset.*

[1] https://www.semanticscholar.org/cord19/download

Three data pre-processing operations are performed per corpus (Papers1, Papers2) in order to create three corpora according to the title, abstract and content:

- Title represents the corpus that contains only paper titles;
- Abstract represents the corpus that contains only paper abstracts;
- Content represents the corpus that contains only paper contents.

We named them PapersX-title, PapersX-abstract and PapersX-content, respectively. See Table 1 for further details and Table 2 for the acronym definitions.

	$NB_d(C)$	$NB_M(d)$	$std(c)$
Papers1			
Papers1-title	9315	15	± 8
Papers1-abstract	9315	180	± 94
Papers1-content	9315	4639	± 359
Papers2			
Papers2-title	32322	13	± 10
Papers2-abstract	32322	168	± 88
Papers2-content	32322	4913	± 720

Table 1 Statistics related to the COVID19-MOOD-data dataset

Abbreviations	Description
$NB_d(C)$	number of documents in the corpus
$NB_M(d)$	average number of words of a document in the corpus
$std(c)$	corpus standard deviation
NN	noun
$NNNN$	matches singular and plural noun terms
JJ	adjective
NP	proper noun

Table 2 Table legend

4 Methodology

Here we outline two complementary term extraction and analysis approaches: the free term extraction approach and the driven term extraction approach. The first one is based on a combination of the type of corpus and the statistical measures, while the second is based on a combination of the type of corpus and the morphosyntactic variation rules.

4.1 The free term extraction approach

The free term extraction approach seeks to ensure that users will be able to extract significant terms related to a specific domain from a given corpus. As we mentioned in section 2, existing tools have been proposed for term and concept extraction. We opted for the BioTex tool to support the free term extraction mode for several reasons:

- BioTex was initially built for medical domain term extraction.
- BioTex uses hybrid measures (linguistic and several statistical measures) for the term extraction process.
- Most existing tools (e.g. *Maui-indexer*[2], *Topia Termextract*[3], *KEA*[4], etc.) are designed for keyword extraction within single documents, and they only function for English language documents, while BioTex is tailored for terminology extraction and supports sets of documents (corpora) and multi-language use.

Three essential parameters related to the BioTex tool are defined below:

- a corpus: this is the data source from which terms are extracted;
- a statistical measure: as mentioned above, the BioText processing approach is based on linguistic and statistical measures. The linguistic parameter is defined by default, but the user must define the statistical parameter, as several exist, in order to run the term extraction process;
- the number of words to be extracted per concept: so called n-grams, this concerns the length of the extracted terms and ranges from 1 to 4_grams for BioTex.

In addition to these parameters, there is the number of linguistic patterns (like NN NN, JJ NP NP, NN NP NP, etc.) that can be associated, but this is preset at 20 by default in BioTex. BioTex also includes patterns for verb terms, such as: NN VBD NN NN, NP NN VBD NN NP, etc. Figure 1 outlines the overall three-step process for free term extraction.

At the end of the BioTex process, extracted terms are classified in two sets: TermSet, which only contains single word terms (SWT), and MultiTermSet, which contains multi-word terms (MWT). By using the Driven

[2] `https://code.google.com/p/maui-indexer/`
[3] `https://pypi.python.org/pypi/topia.termextract/1.1.0`
[4] `http://www.nzdl.org/Kea/`

Extraction process (with FASTR), we can capture the entire term for a given incomplete one obtained during the first step (Free Extraction). The Driven Extraction process step uses incomplete terms to capture the entire terms in the document. For example, if "higher risk acute" or "higher risk area" terms are extracted in the Free Extraction process step, an entire term which could be "higher risk acute care area" will be obtained during the Driven Extraction process.

4.2 The driven term extraction approach

This extraction approach seeks to ensure that the terms extracted using BioTex could be used to improve the domain terminology. From a given term, the process aims to extract some variations of this term that exist in the corpus.

The overall processing under this approach is handled with FASTR (Jacquemin, 1994). FASTR is a rule-based linguistic tool that generates morphosyntactic variants of terms. We respectively note *NN, NNS, NNP, NNPS* for noun paterns, *VB, VBD, VBG, VBN, VBP, VBZ* for verbs, *RB, RBR, RBS* for adverbs and finally *JJ, JJR, JJS* for adjectives. It enables extraction of variants of a given term in full-text documents. For a given term, FASTR helps extract nearby or long terms that contain the initial term. Figure 1 illustrates the two steps (4 and 5) of the driven term extraction approach. For a given term, FASTR helps extract nearby or long terms that contain the initial one.

The driven process has the advantage of extracting relevant new terms that BioTex cannot extract from the corpus.

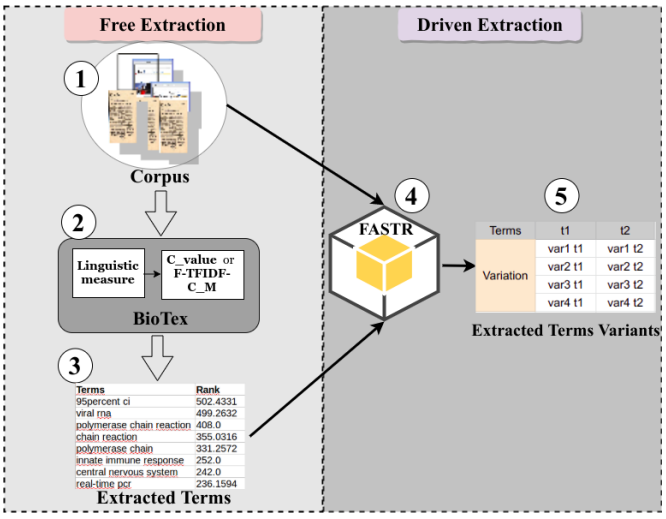

Fig. 1 The Free and Driven process for term extraction using BioTex and FASTR

4.3 Proposed combination for term extraction

Based on the elements given in sections 4.1 and 4.2, we propose a workflow in Figure 2 for term extraction and analysis dedicated to scientific papers. We outline this workflow according to the type of corpus, measure, and approach:

- *The type of corpus*: as described in the data section, for a given paper, we considered three parts to build the corresponding corpora, i.e. the Title (T), Abstract (A) and Content (C);
- *The measures*: BioTex integrates several statistical measures, each of which uses a specific strategy to compute the term score. In this case, we selected the two measures C_Value and F-TFIDF-C_M. C_Value indicates the importance of terms that appear most frequently in a document, based on the idea that the frequency of appearance of a term in the document reflects its importance in the document. Moreover, based on frequency criteria, C_Value favors multi-word term extraction by taking into account nested terms (e.g. virus) in multi-word terms (e.g. influenza virus) (Frantzi et al., 2000). F-TFIDF-C_M represents the harmonic mean of the two C_Value and TF-IDF values, which ranks terms by weight according to their relevance in the document while taking the whole corpus into account (Lossio-Ventura et al., 2016). C_Value and F-TFIDF-C_M are complementary, as the first favors relevant MWT extraction while the second gives weight to discriminant terms.
  For each measure, the aim is to organise the extracted terms in to five sets. 1) terms corresponding to the Title corpus Set(T), 2) terms corresponding to the Abstract corpus Set(A), 3) terms corresponding to the Content corpus Set(C), 4) terms that intersect within the Title and the Abstract corpus Set(TA), and 5) terms that intersect within the Title and the Content corpus Set(TC).
- *The approach*: terms could be extracted using both a given corpus and a specific statistical measure in a free extraction approach. Moreover, for the driven process, term variations are extracted by using both a given corpus and specific set of terms. The set of terms could be defined from the output of the previous approach.

5 Experiments

To set the parameters, throughout our study we used C_Value and F-TFIDF-C_M as statistical measures, 50

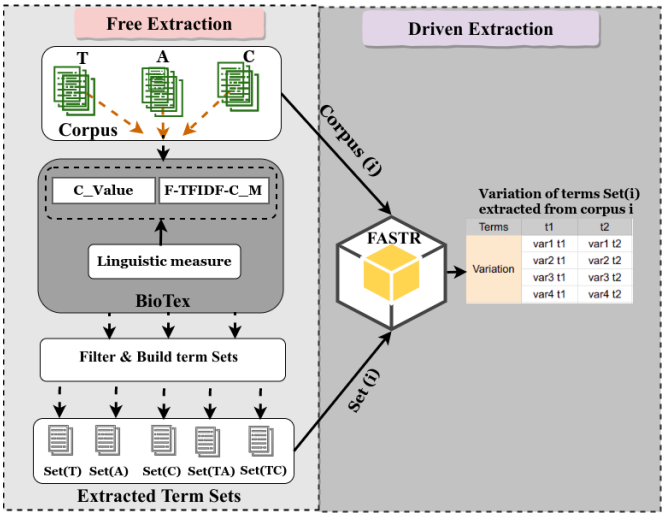

Fig. 2 Proposed combination for term extraction

different patterns or term extraction rules, and a number of words ranging from 1 to 4-grams ($n = 1, 2, 3, 4$). These parameters are applied for corpora described in section 3. The choice of C_Value and F-TFIDF-C_M is based on the findings of previous studies (Lossio-Ventura et al., 2016; Frantzi et al., 2000) which showed that both allow efficient SWT and MWT extraction.

Before applying BioTex, some specific pre-processes were applied for the Papers1-content and Papers2-content corpora due to their size. Papers1-content was divided into 09 sub-corpora (8 corpora of 1000 documents each and 1 corpus of 1315 documents) and Papers2-content into 32 sub-corpora (31 corpora of 1000 documents each, and 1 corpus of 1332 documents). Each corpus was partitioned into smaller units to enhance scalability. The results obtained from the smaller units were then composed by computing the average ranked values. The final rank for a given term was thus equal to the average of its ranked values in all sub-corpora in which it was present. The final result gave a set of terms, listed in ascending order according to the ranking values.

Table 3 shows an example of the MWT set obtained using BioTex. The *Terms* column contains the extracted terms, the *in_umls* column indicates if the corresponding term is available in the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) or not, and *rank* shows the significance of the term based on statistical measures in the whole list of terms for a given corpus. In our study, we used the UMLS Metathesaurus as reference for the extracted terms as our study is linked to a biomedical terminology analysis. This comparison aimed to separate new terminologies or terminologies that were not yet listed in the Metathesaurus.

Terms	in_umls	rank
public health	1	1602.3971
respiratory syndrome	0	1481.9399
infectious disease	1	1198.2317
virus infection	1	1126.9083
influenza virus	1	1023.8858
immune response	1	1008.0362

Table 3 Example of BioTex ouput

5.1 The free term extraction approach

We used BioTex, as outlined in section 4.1, to extract terms from corpora in free mode. Several analyses are performed below on the obtained results. To this end, we conducted the experiments to address three main questions: 1) for each corpus, what are the most representative terms or domain concepts (terms that summarize the main content of the corpus) per statistical measure? 2) for each corpus, what are the most representative concepts for both measures? and 3) what are the discriminant and common concepts of the overall corpus?

For each case, we determined if the extracted terms exist or not in the UMLS Metathesaurus.

5.1.1 Corpus representative terms

In this section, we illustrate how representative terms can be extracted from different datasets. Based on the BioTex ranking measures, a term is more important than another one in a given corpus if it has a higher ranking than the other term.

Figure 3 shows representative terms for the Title, Abstract and Content corpora with the corresponding statistical measures (see Table 7 and 8 for more details).

This figure highlights which terms are important in each part of the Papers. Note that the extracted terms are different for each measure and sub-corpus, but some of them are similar for both. For example, terms like *public health, immune responses* are extracted using both measures from the Abstract corpus.

In order to quantitatively display the number of representative intersecting terms from different corpora, we show common terms between Title vs Abstract, and Title vs Content corpora for the Papers2 corpus in Figure 4. For both measures, Title terms are more representative in the Abstract than in the Content of Papers, i.e. 57% and 27% compared to 28% and 5%, respectively, for Title vs Abstract and Title vs Content. However, we noted that terms extracted with C_Value generated more common terms than those extracted with F-TFIDF-C_M. The common terms represent terms extracted at once in the Title, Abstract and Content cor-

pus for each measure.

As indicated, extracted terms were compared with the UMLS Metathesaurus. Table 4 shows the TOP@20 terms extracted for the Papers1-content corpus using C_Value and F-TFIDF-C_M measures. Bold terms are not in the UMLS Metathesaurus.

According to these TOP@20 terms, we can see that:

- the majority of the SWTs are in the UMLS Metathesaurus for both statistical measures (C_Value or F-TFIDF-C_M);
- for MWTs, several terms are not in the UMLS Metathesaurus. These terms can be categorized as:
  - *UMLS sub-terms:* these are terms that do not exactly match to those present in the UMLS Metathesaurus but could be part of them. For example, *health emergency* is part of terms like *Emergency Health Services* in the UMLS Metathesaurus;
  - *New terms:* these terms are not in the UMLS Metathesaurus, but are meaningful (or not) in the COVID-19 context. For example, terms like *close contact* relate to the COVID-19 contagion mode.

Figures 5 and 6 illustrate the number of terms out of the TOP@100 terms (in percentage) for each measure (C_Value, F-TFIDF-C_M) and dataset (Papers1-title, Papers2-title):

- In_UMLS: the number of terms in the UMLS Metathesaurus;
- Not_In_UMLS_V: the number of terms that do not exactly match the UMLS terms, but have some variants or are part of the UMLS terms;
- Not_In_UMLS: the number of terms that do not match the UMLS terms at all. We indicate these as new terms. Terms which are not in the UMLS Metathesaurus but which could have greater meaning in the study context or which could be added to the UMLS Metathesaurus.

According to these statistics, we first note that the C_Value and F-TFIDF-C_M measures enable extraction of more conventional terms or terms in the UMLS Metathesaurus regardless of the corpus. Secondly, we note that F-TFIDF-C_M generates more new terms (Not In UMLS) than C_Value regardless of the corpus. Finally, the number of new terms is more substantial with MWTs (figure 6) than SWTs (figure 5) regardless of the measure.

5.1.2 Relevant term extraction from corpora for both measures

This involves quantitative and qualitative analysis of the terms extracted within each corpus, while taking both measures (C_Value and F-TFIDF-C_M) into account. In other words, it consists of analysing terms obtained for both measures, i.e. terms detected at the same time, and also terms specific to each of them.

The quantitative analysis aims to highlight, for each dataset, the number of terms obtained by each measure, the number of terms obtained for both measures, and which are available or not in the UMLS Metathesaurus. While the qualitative measure aims to highlight, in each case, how the terms obtained are important or not regarding the study domain.

For the data representation, we take advantages of Venn Diagram (Ho et al., 2021), see in Appendix-Figure 10 the distribution of the Papers2-title corpus terms. Terms are organised in different sections. For example, *gene expression, human coronavirus, case report, public health, respiratory syncytial virus, etc.* are available in UMLS Metathesaurus and are recognized by both measures (C_Value and F-TFIDF-C_M). According to the study domain, these terms will tend to be more representative and important in the whole corpus. Moreover, for each measure there are new terms which are not in the UMLS Metathesaurus.

5.1.3 Discriminant and common term extraction from corpora

In this case, term analysis is performed per dataset or by jointly considering multiple corpora, i.e. between Title, Abstract and Content corpora. Appendix-Figure 11 corresponds to discriminant and common term extraction from Papers1-title, Papers1-abstract and Papers1-content.
There are common terms in the overall corpus such as *gene expression, virus replication, influenza virus, etc.*. These terms tend to be relevant in the Title, Content and Abstract corpora.

Moreover, *[respiratory infection, acute respiratory infection, etc.]*, *[innate immune response, endoplasmic reticulum, etc.]*, and *[nucleotide sequences, room temperature, etc.]* are discriminant terms in the Title, Abstract and Content corpora.

5.2 The driven term extraction process

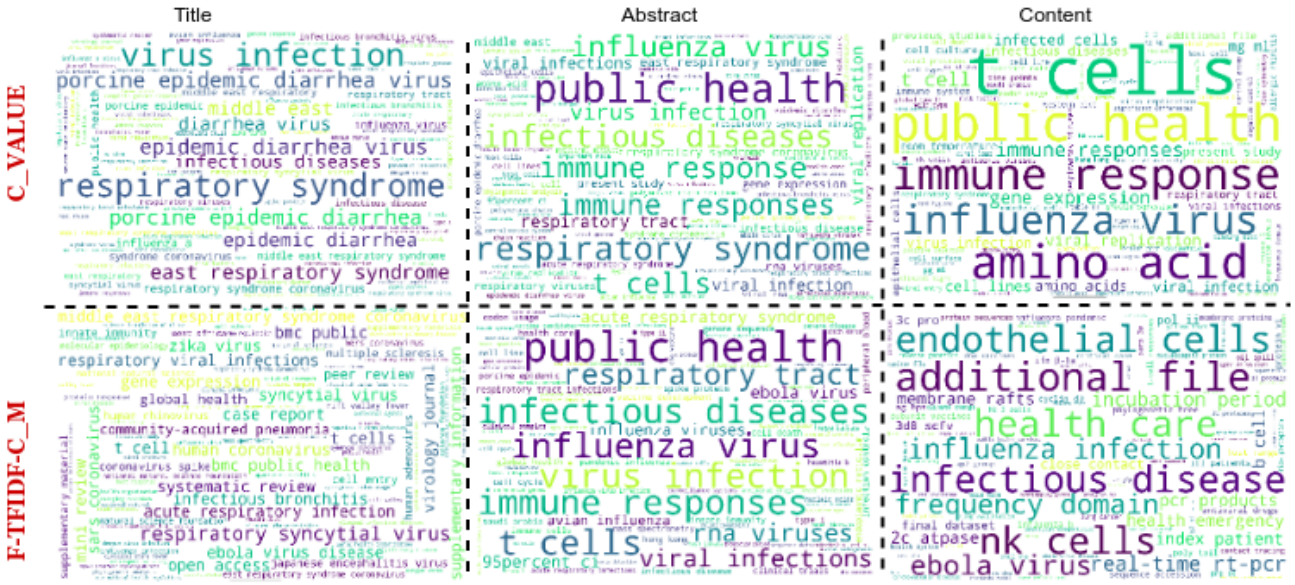

Fig. 3 Representative terms from Papers1

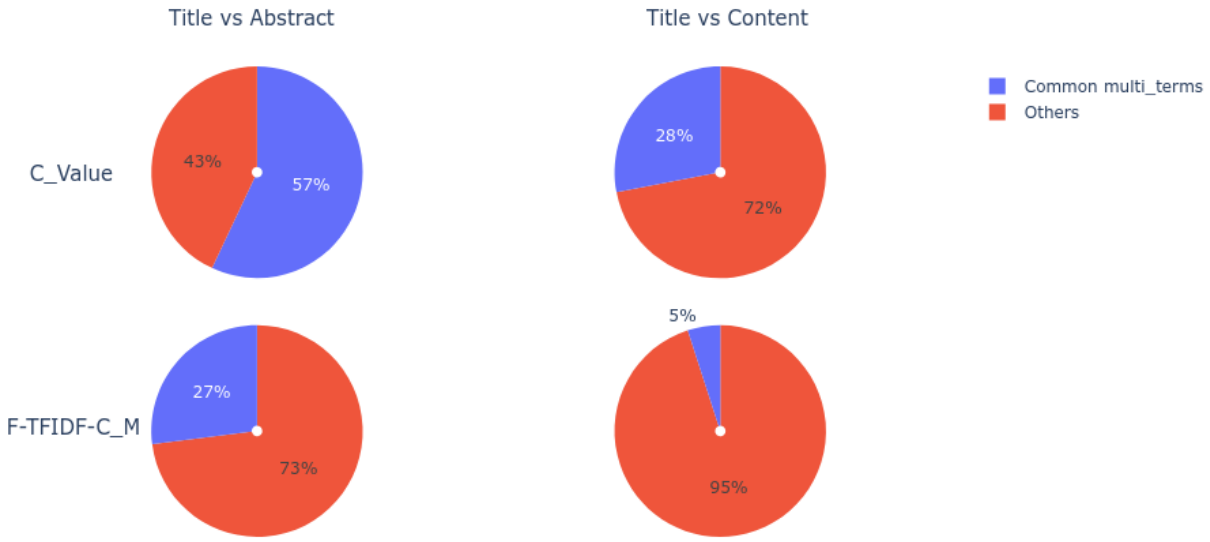

Fig. 4 Common terms in Papers2

We performed a driven term extraction strategy using FASTR. Our proposal addresses two main questions: 1) For a given set of terms, how can new and relevant terms variants be extracted from a corpus based on the terms? 2) Do some of the new terms exist in the UMLS Metathesaurus? In our experiment, we used the common terms extracted in section 5.1.3 based on the fact that they were more representative and relevant throughout the corpora.

Figure 7 shows an example of variant terms extracted with the term *infectious disease*. Among these variants, we only show those which are not in the UMLS Metathesaurus since they are new and might be more informative.

Table 5 contains a list of TOP@10 variants extracted with six initial terms. Among them, we highlighted (in bold) terms matching terms in the UMLS Metathe-

C_Value Measure					
	SWTs				
TOP 20	cells	virus	infection	protein	study
	data	**figure**	al	patients	expression
	rna	analysis	result	disease	p
	mice	c	**samples**	influenza	number
	MWTs				
TOP 20	t cells	public health	amino acid	immune response	gene expression
	viral replication	**infected cells**	cell lines	viral infection	virus infection
	mg ml	infectious diseases	**present study**	respiratory tract	epithelial cells
	previous studies	room temperature	cell culture	**additional file**	viral infection

F-TFIDF-C_M Measure					
	SWTs				
TOP 20	mice	patients	influenza	proteins	health
	dna	vaccine	transmission	research	model
	children	outbreak	vaccination	e	china
	peptide	fusion	network	percent	mers-cov
	MWTs				
TOP 20	**additional file**	infectious disease	nk cells	health care	endothelial cells
	frequency domain	ebola virus	**influenza infection**	**real-time rt-pcr**	incubation period
	health emergency	**index patient**	**membrane rafts**	**pcr products**	**2c atpase**
	b cell	**close contact**	**final dataset**	**3d8 scfv**	pol ii

In bold terms not in the UMLS thesaurus

Table 4 TOP@20 terms extracted from Paper1-content using C_Value and F-TFIDF-C_M - SWTs vs MWTs

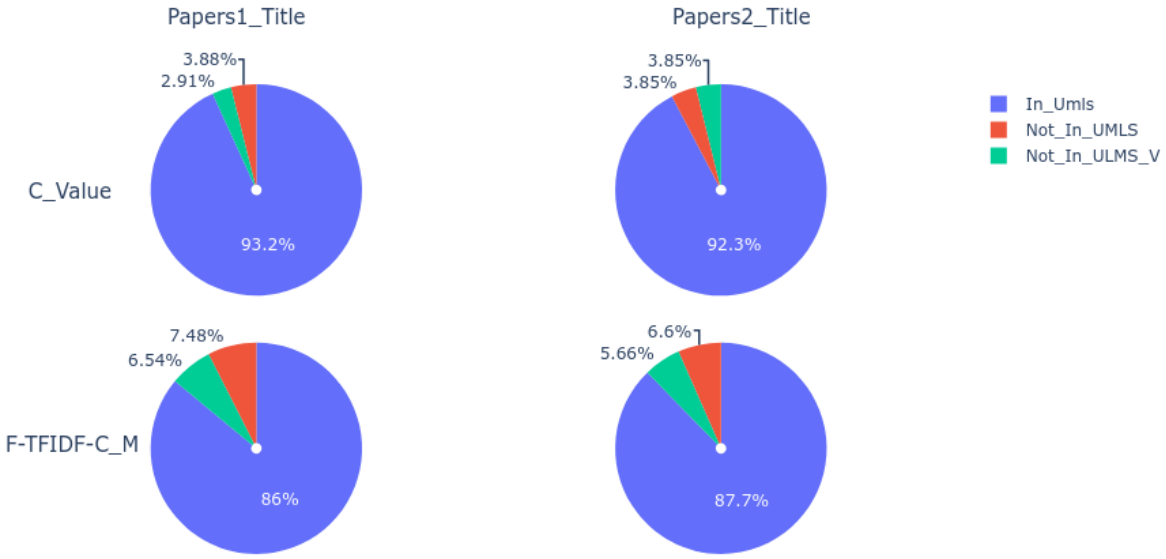

Fig. 5 C_Value vs F-TFIDF-C_M SWTs

saurus. Like free mode extraction, term variant mode may be used to extract useful terms.

5.3 Combined strategies for term analysis

Combined strategies for term analysis concern two levels: 1) Intra-corpus term extraction, and 2) Inter-corpus term extraction.

– Combined intra-corpus term extraction strategies: these are geared towards extracting common or dis-

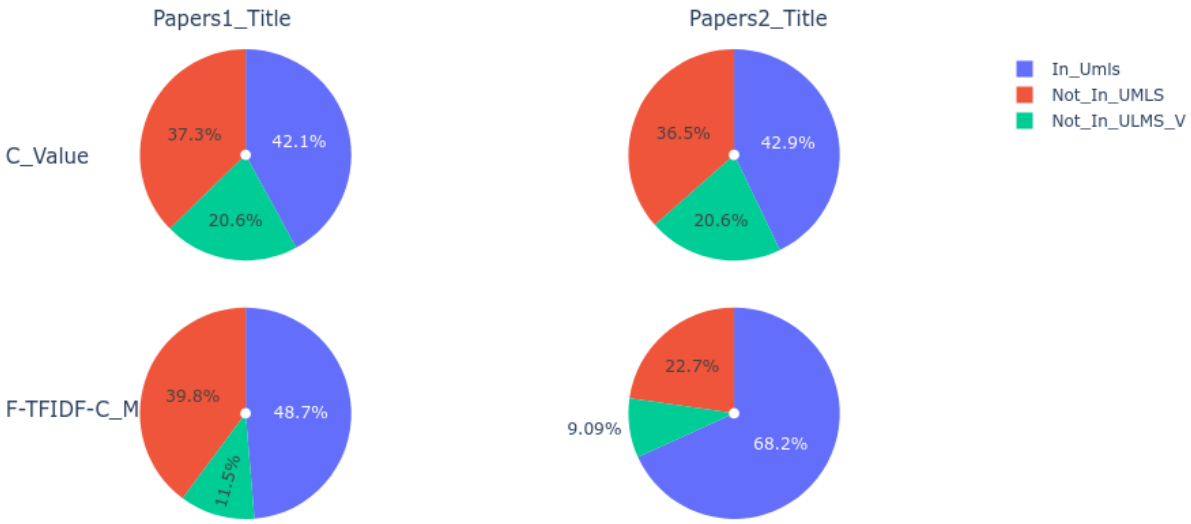

Fig. 6 C_Value vs F-TFIDF-C_M MWTs

Terms	infectious disease	virus replication	laboratory tests	respiratory syndrome	preventive measure	syndrome coronavirus
Variations	diseases including infectious	**replication competent viruses**	laboratory confirmation tests	**respiratory distress syndrome**	**preventive measures**	syndrome coronavirus-related coronavirus
	infectious pulmonary diseases	replication of N1347A virus	**laboratory testing**	respiratory acute syndrome	preventive hygienic measures	Syndrome human coronavirus
	infectious bursal disease	virus optimal replication	Testing presents isolation laboratories	syndrome coronavirus and respiratory	prevention community-engaged measures	Syndromic Surveillance Coronavirus
	infectious lung diseases	replicating influenza viruses	laboratory diagnostic testing	respiratory tract syndromic	preventive health measures	syndrome virus coronavirus
	infectious acute disease	replication of human viruses	laboratory genomic testing	respiratory insufficiency syndrome	preventive behavioral measures	Coronavirus Associated Syndromes

Terms in the UMLS Metathesaurus in bold

Table 5 Term extraction variations using FASTR

criminant terms from a given corpus. To this end, extracted terms from both measures are compared. We show the process in Figure 8, where the set of terms Set(Cp) extracted from the corpus Cp (Title, Abstract or Content) using each measure (C_Value, F-TFIDF-C_M) are jointly compared with the UMLS Metathesaurus terms. Set A represents corpus terms specifically extracted with C_Value, set B represents terms that are specific to F-TFIDF-C_M, while set C represents common terms from both measures and UMLS Metathesaurus elements. We consider that sets A and B are discriminant terms of the corpus according to the measures, and otherwise set C is considered as containing common terms or the most representative terms of the corpus.

The new term extraction process with FASTR is run with one of the combined sets (discriminant or common) and the corpus.

– Combined inter-corpus term extraction strategies: these are geared towards extracting common and discriminant terms, while taking several corpora into account for a given measure. As illustrated in Fig-

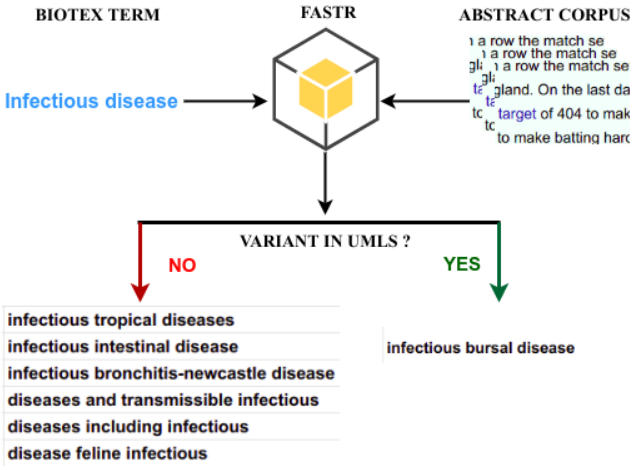

Fig. 7 Example of term variants

ure 9, for each measure (C_Value or F-TFIDF-C_M), the sets of terms Set(Cp1), Set(Cp2), Set(Cp3) are extracted respectively from corpus Cp1 (Title), Cp2 (Abstract), and Cp3 (Content). These sets are compared in order to compute the common term set D for both corpora, and discriminant term sets A, B, C, respectively, for corpora Cp2, Cp1 and Cp3. In this context, new terms are extracted using one of the combined sets with one corpus (Cpx).

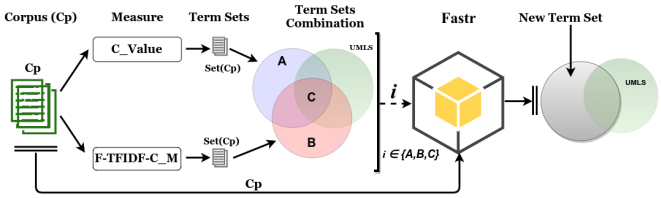

Fig. 8 Combined intra-corpus term extraction strategies

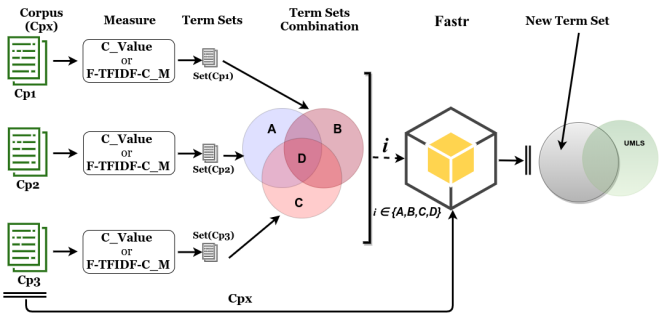

Fig. 9 Combined inter-corpus term extraction strategies

6 Case study: Epidemic intelligence

Epidemic intelligence (EI) aims to detect, investigate and monitor potential health threats in a timely manner (Paquet et al., 2006). In addition to conventional surveillance system monitoring, such as outbreak notifications from the World Organisation for Animal Health (OIE), the EI process increasingly mainstreams unstructured data from informal sources such as online news. Several web-based surveillance systems have been developed and used to support public health and animal health surveillance (ProMED (Madoff, 2004), HealthMap (Freifeld et al., 2008), GPHIN (Mykhalovskiy and Weir, 2006), PADI-web (Valentin et al., 2020a), etc.). In this case study, we focused on the choice keywords with the PADI-Web system for COVID-19 surveillance (i.e. driven surveillance) and for monitoring unknown diseases (i.e. syndromic surveillance).

The Platform for Automated extraction of Disease Information from the web (PADI-web[5]) is an automated surveillance system for monitoring the emergence of animal infectious diseases, including zoonoses (Arsevska et al., 2018; Valentin et al., 2020a). PADI-web monitors Google News through specific really simple syndication (RSS) feeds, targeting diseases of interest (e.g. African swine fever, avian influenza, etc.). PADI-web also uses unspecific RSS feeds, consisting of combinations of symptoms and hosts (i.e. species), thus allowing syndromic surveillance and detection of unusual disease events. RSS feeds consists of combinations of different categories of terms (i.e. keywords) including symptoms, disease names and species.

PADI-Web has been used for monitoring COVID-19 disease (Valentin et al., 2020b). In this context, the choice of COVID-19 surveillance terms is crucial.

In the following subsections, we discuss the choice of terms given by ITEXT-BIO to use in the PADI-Web system (Valentin et al., 2020a) and other web-based surveillance systems (Madoff, 2004; Freifeld et al., 2008; Mykhalovskiy and Weir, 2006) for COVID-19 and syndromic surveillance. This enables evaluation of the relevance of terms generated by our approach for a dedicated task, i.e. web-based health surveillance.

6.1 Relevant term extraction

We compared the relevance of the top 10 terms extracted from Papers2 corpora with either C_Value or

[5] https://padi-web.cirad.fr/en/

F-TFIDF-C (Table 6). Table 9 gives more details on these terms. The relevance was assessed by classifying the terms in one or more of the following categories:

- COVID-19 surveillance: epidemiological terms specific to COVID-19 (e.g. *coronavirus spike*).
- Syndromic surveillance: epidemiological terms not specific to a particular disease (e.g. *infectious bronchitis*).
- Domain relevant: terms related to health, i.e. either to specific diseases (e.g. *porcine epidemic diarrhoea*) or unspecific (e.g. *virus infections*). The Domain relevant category thus includes the two previous categories, plus diseases other than COVID-19.
- Part of disease multiword expression (MWE): part of a multiword expression corresponding to a disease name (e.g. *East respiratory syndrome* for *Middle East syndrome coronavirus*).

Among the terms extracted with C_Value from Titles, Abstracts or Titles and Abstracts, six to seven were parts of disease MWE. Only one term extracted with F-TFIDF-C_M was a part of disease MWE. C_Value could thus be of particular interest for extracting disease name variants, even if they are incomplete. For domain relevant COVID-19 surveillance and syndromic surveillance terms, F-TFIDF-C_M obtained better results than C_Value, even when the frequency of relevant terms was low (from one to five out of ten terms). No common terms were extracted from (Title + Abstract) or from (Title + Content) using F-TFIDF-C_M. Using C_Value, only three common terms were extracted from Title + Content. Among the top 10 terms extracted from Title + Abstract with these metrics, seven were parts of disease MWE. Regardless of the term category, we extracted more relevant terms from Titles and Abstracts than from Contents. This is in line with the fact that Title and Abstracts are more rich in key information and relevant terms due to their length limitation.

6.2 Driven term extraction

We selected terms extracted in section 6.1: *respiratory tract, viral infections, SARS coronavirus, incubation period, influenza virus, respiratory infections and infectious diseases*. We randomly extracted the variants with FASTR (section 4.2). An epidemiologist manually evaluated the relevance of 10 randomly selected variants per term. Among the 60 evaluated terms (see Table 10), 72% (43/60) were relevant and 7% (4/60) were irrelevant. For 13 variants (22%), the relevance could not be assessed because the expression was truncated and ambiguous, such as "disease has an infectious" for the term "infectious diseases". FASTR thus seems to be an effective tool for generating term variants efficiently. However, we noted that FASTR generated up to 774 variants for a single term. Thus, to avoid random selection of terms, it would be interesting to compute a relevance index that could be used to rank the proposed variants. Besides, several extracted variants were fragments of expressions that could not be evaluated. This issue could be overcome by displaying the variant context (i.e. the sentence in which the variants appeared).

7 Conclusion

In this paper we describe ITEXT-BIO, a generic methodology for biomedical term extraction. We show how it allows users to extract terms (or concepts) from different types of textual data using several combined strategies. The free term extraction approach extracts terms from corpora, while the driven term extraction approach extracts, from a corpus and a set of terms, a set of variations of these terms.

We illustrate that the proposed combined strategies based on statistical measures and textual segments help efficiently extract and categorize terms (representative, discriminant and new terms) from a corpus or corpora. We also quantitatively and qualitatively analysed the extracted terms to determine those related to the study domain and those that could be considered as emerging terminology for disease monitoring.

Our future studies will focus on term extraction and analysis by: (i) taking different sections of papers into account and applying the methodology to different types of corpora derived from newspapers or social media such as Twitter, (ii) considering combinations of tools other than BioTex, and (iii) introducing word embedding strategies like BERT (Devlin et al., 2018) to capture semantic aspects of the extracted terms in order to reduce context ambiguity.

Acknowledgements This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD 003. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. This study was partially funded by the French Agricultural Research Centre for International Development (CIRAD) the French General Directorate for Food (DGAL), and the SONGES Project (FEDER and Occitanie Region). The research was also supported by the French National Research Agency (ANR) under the Investments for the Future Program, referred to as ANR-16-CONV-0004, #DigitAg.

Corpus (Papers2)	Measure	n	Domain relevant	COVID-19 surveillance	Syndromic surveillance	Part of disease MWE
title	C_Value	10	3	0	2	6
title	F-TFIDF-C_M	9	4	1	1	1
abstract	C_Value	10	1	0	0	6
abstract	F-TFIDF-C_M	10	5	1	2	1
content	C_Value	10	0	0	0	1
content	F-TFIDF-C_M	10	2	0	2	0
title + abstract	C_Value	10	3	0	0	7
title + abstract	F-TFIDF-C_M	0	-	-	-	-
title + content	C_Value	3	1	0	0	2
title + content	F-TFIDF-C_M	0	-	-	-	-

Table 6 Relevance of terms extracted from Papers2 depending on the metrics (C_Value or F-TFIDF-C_M)

References

Arsevska E, Valentin S, Rabatel J, de Goër de Hervé J, Falala S, Lancelot R, Roche M (2018) Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLOS ONE 13(8):e0199960, DOI 10.1371/journal.pone.0199960, URL https://dx.plos.org/10.1371/journal.pone.0199960

Azarafza M, Feizi-Derakhshi MR, Shendi MB (2020) Textrank-based microblogs keyword extraction method for persian language. Conference: 3rd International Congress on Science and EngineeringAt: Hamburg - Germany

Bodenreider O (2004) The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 32(suppl_1):D267–D270

Bracewell DB, Ren F, Kuriowa S (2005) Multilingual single document keyword extraction for information retrieval. In: 2005 International Conference on Natural Language Processing and Knowledge Engineering, IEEE, pp 517–522

Brill E (1992) A simple rule-based part of speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing, Association for Computational Linguistics, USA, ANLC '92, p 152–155, DOI 10.3115/974499.974526, URL https://doi.org/10.3115/974499.974526

Campillos Llanos L, Sandoval AM, Guirao J (2013) An automatic term extractor for biomedical terms in spanish. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013)

Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A (2020) Yake! keyword extraction from single documents using multiple local features. Information Sciences 509:257–289, DOI 10.1016/j.ins.2019.09.013

Conrado M, Pardo T, Rezende SO (2013) A machine learning approach to automatic term extraction using a rich feature set. In: Proceedings of the 2013 NAACL HLT Student Research Workshop, pp 16–23

Cram D, Daille B (2016) Terminology extraction with term variant detection. In: Proceedings of ACL-2016 System Demonstrations, pp 13–18

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Duari S, Bhatnagar V (2020) Complex network based supervised keyword extractor. Expert Systems with Applications 140:112876

Foo J (2009) Term extraction using machine learning. Linkoping University, LINKOPING

Frantzi K, Ananiadou S, Mima H (2000) Automatic recognition of multi-word terms:. the c-value/nc-value method. International journal on digital libraries 3(2):115–130

Freifeld CC, Mandl KD, Reis BY, Brownstein JS (2008) HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. Journal of the American Medical Informatics Association 15(2):150–157, DOI 10.1197/jamia.M2544, URL https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M2544

Habibi M, Popescu-Belis A (2015) Keyword extraction and clustering for document recommendation in conversations. IEEE/ACM Transactions on audio, speech, and language processing 23(4):746–759

Ho SY, Tan S, Sze CC, Wong L, Goh WWB (2021) What can venn diagrams teach us about doing data science better? International Journal of Data Science and Analytics 11(1):1–10

Jacquemin C (1994) Fastr: A unification-based front-end to automatic indexing. In: Intelligent Multimedia Information Retrieval Systems and Management - Volume 1, Le centre de hautes études internationales d'informatique documentaire, Paris, FRA, RIAO '94, p 34–47

Ji L, Sum M, Lu Q, Li W, Chen Y (2007) Chinese terminology extraction using window-based contextual information. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp 62–74

Joung J, Kim K (2017) Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technological Forecasting and Social Change 114:281–292

Kageura K, Umino B (1996) Methods of automatic term recognition: A review. Terminology International Journal of Theoretical and Applied Issues in Specialized Communication 3(2):259–289

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2014a) Biomedical terminology extraction: A new combination of statistical and web mining approaches. In: JADT: Journées d'Analyse statistique des Données Textuelles, pp 421–432

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2014b) Biotex: A system for biomedical terminology extraction, ranking, and validation. In: Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272, CEUR-WS.org, Aachen, DEU, ISWC-PD'14, p 157–160

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2014c) Yet another ranking function for automatic multiword term extraction. In: International Conference on Natural Language Processing, Springer, pp 52–64

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2016) Biomedical term extraction: overview and a new methodology. Information Retrieval Journal 19(1-2):59–99

Madoff LC (2004) ProMED-mail: an early warning system for emerging diseases. Clinical infectious diseases 39(2):227–232, URL https://academic.oup.com/cid/article-abstract/39/2/227/327615

Matsuo Y, Ishizuka M (2004) Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13(01):157–169

Maynard D, Yankova M, Kourakis A, Kokossis A (2005) Ontology-based information extraction for market monitoring and technology watch. In: ESWC Workshop" End User Apects of the Semantic Web"), Heraklion, Crete

Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF (2008) Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics 17(01):128–144

Mykhalovskiy E, Weir L (2006) The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Canadian journal of public health = Revue canadienne de sante publique 97(1):42–44, DOI 10.17269/cjph.97.756

Neifar W, Hamon T, Zweigenbaum P, Khemakhem ME, Belguith LH (2016) Adaptation of a term extractor to arabic specialised texts: First experiments and limits. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp 242–253

Oliver A, Vàzquez M (2015) Tbxtools: a free, fast and flexible tool for automatic terminology extraction. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, pp 473–479

Pais V, Ion R (2020) Termeval 2020: Racai's automatic term extraction system. In: COMPUTERM

Paquet C, Coulombier D, Kaiser R, Ciotti M (2006) Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Eurosurveillance 11(12):5–6, DOI 10.2807/esm.11.12.00665-en, URL https://www.eurosurveillance.org/content/\10.2807/esm.11.12.00665-en

Pazienza MT, Pennacchiotti M, Zanzotto FM (2005) Terminology extraction: an analysis of linguistic and statistical approaches. In: Knowledge mining, Springer, pp 255–279

Ramos J, et al. (2003) Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, New Jersey, USA, vol 242, pp 133–142

Rigouts Terryn A, Hoste V, Drouin P, Lefever E (2020) Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In: 6th International Workshop on Computational Terminology (COMPUTERM 2020), European Language Resources Association (ELRA), pp 85–94

Shah H, Khan MU, Fränti P (2019) H-rank: a keywords extraction method from web pages using pos tags. In: 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), IEEE, vol 1, pp 264–269

Valentin S, Arsevska E, Falala S, de Goër J, Lancelot R, Mercier A, Rabatel J, Roche M (2020a) PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. Computers and Electronics in Agriculture 169:105163, DOI 10.1016/j.compag.2019.105163, URL http://www.sciencedirect.com/science/article/pii/\S0168169919310646

Valentin S, Mercier A, Lancelot R, Roche M, Arsevska E (2020b) Monitoring online media reports for early detection of unknown diseases: Insight from a retro-

spective study of covid-19 emergence. Transboundary and emerging diseases

Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney RM, Liu Z, Merrill W, Mooney P, Murdick DA, Rishi D, Sheehan J, Shen Z, Stilson B, Wade AD, Wang K, Wilhelm C, Xie B, Raymond DM, Weld DS, Etzioni O, Kohlmeier S (2020a) Cord-19: The covid-19 open research dataset. ArXiv abs/2004.10706

Wang R, Liu W, McDonald C (2016) Featureless domain-specific term extraction with minimal labelled data. In: Proceedings of the Australasian Language Technology Association Workshop 2016, pp 103–112

Wang X, Zhang L, Klabjan D (2020b) Keyword-based topic modeling and keyword selection. arXiv preprint arXiv:200107866

Whissell JS, Clarke CL (2011) Improving document clustering using okapi bm25 feature weighting. Information retrieval 14(5):466–487

Yao Xm, GAN Jh, Jian X (2017) Concept extraction based on hybrid approach combined with semantic analysis. DEStech Transactions on Engineering and Technology Research

Appendices

A TABLES

F-TFIDF-C_M

Title Corpus		Abstract Corpus		Content Corpus	
terms	**rank**	**terms**	**rank**	**terms**	**rank**
respiratory syncytial virus	1.9880	public health	1.9986	additional file	1.9976
middle east respiratory syndrome coronavirus	1.9846	infectious diseases	1.9979	infectious disease	1.997
systematic review	1.9842	immune responses	1.9976	nk cells	1.997
open access	1.9819	influenza virus	1.9976	health care	1.996
zika virus	1.9819	t cells	1.9975	endothelial cells	1.9957
gene expression	1.9795	virus infection	1.9974	frequency domain	1.9957
virology journal	1.9788	respiratory tract	1.9973	ebola virus	1.9948
human coronavirus	1.976	viral infections	1.9969	influenza infection	1.9943
case report	1.9756	rna viruses	1.9967	real-time rt-pcr	1.9933
syncytial virus	1.9752	acute respiratory syndrome	1.9961	incubation period	1.99325
t cell	1.9746	95percent ci	1.996	health emergency	1.9932
infectious bronchitis	1.9726	ebola virus	1.9945	index patient	1.9932
sars coronavirus	1.9723	influenza viruses	1.9943	membrane rafts	1.9931
bmc public health	1.9701	avian influenza	1.9939	pcr products	1.9929
t cells	1.9689	respiratory tract infections	1.9938	2c atpase	1.9926
acute respiratory infection	1.9672	health care	1.9925	b cell	1.9924
mini review	1.9636	hepatitis c	1.9922	close contact	1.9924
respiratory viral infections	1.9636	type i	1.9918	final dataset	1.9922
bmc public	1.9625	cell line	1.9914	3d8 scfv	1.9921
ebola virus disease	1.9592	spike protein	1.9909	pol ii	1.992
supplementary information	1.9574	codon usage	1.9908	3c pro	1.992
community-acquired pneumonia	1.9543	pandemic influenza	1.9907	influenza pandemic	1.9919
global health	1.9543	endoplasmic reticulum	1.9904	phylogenetic tree	1.9918
peer review	1.9543	saudi arabia	1.9904	protein vi	1.9917
japanese encephalitis virus	1.9512	innate immunity	1.9903	ag nps	1.9916
innate immunity	1.9488	porcine epidemic	1.9903	influenza b	1.99125
multiple sclerosis	1.9488	global health	1.9902	ifn $\beta - 1a$	1.991
human rhinovirus	1.9466	vaccine development	1.9901	ill patients	1.9908
supplementary material	1.9442	cell death	1.9898	poly tail	1.9908
cell entry	1.9417	infectious disease	1.9896	host range	1.9906
coronavirus spike	1.9417	peripheral blood	1.9895	cyclin d3	1.9903
human adenovirus	1.9417	hong kong	1.9894	sequence accession	1.9903
east respiratory syndrome coronavirus	1.9414	immune cells	1.9888	antiviral drugs	1.9897
mers coronavirus	1.9388	cell cycle	1.9886	subunit vaccines	1.9897
west africa	1.9388	clinical trials	1.9885	protein sequences	1.9895
molecular epidemiology	1.9323	infection control	1.9884	oil spill	1.9895
national natural science	1.931	mass spectrometry	1.9883	swine flu	1.9894
natural science foundation	1.931	genome sequence	1.9881	membrane proteins	1.9893
rift valley fever	1.931	clinical samples	1.9877	contact tracing	1.9891
national natural science foundation	1.9307	acute respiratory infections	1.9874	sars 3a	1.9889
influenza infection	1.9284	severe disease	1.9868	critical care	1.9888
protein response	1.9284	hepatitis b	1.9864	hk-2 cells	1.9888
science foundation	1.9284	host response	1.9864	ap2 group	1.9887
supplementary materials	1.9284	type ii	1.9864	prp sc	1.9887
natural science	1.9241	nucleic acids	1.9862	t-cell responses	1.9887
respiratory syndrome coronavirus infection	1.9241	surveillance systems	1.9859	dna vaccines	1.9886
influenza virus	1.9212	influenza virus infection	1.9852	reverse genetics	1.9886
obstructive pulmonary disease	1.92	antiviral drugs	1.9851	health system	1.9884
emerging microbes	1.9193	dna vaccine	1.9847	$b7 - h1$	1.9884
original research	1.9193	influenza infection	1.9845	hcv infection	1.9883
retrospective study	1.9193	reference genes	1.9842	lung cancer	1.9879
phylogenetic analysis	1.9153	cell types	1.984	nucleocapsid protein	1.9879
respiratory syndrome coronavirus	1.9151	b cell	1.9835	3c protease	1.9878
clinical characteristics	1.9138	vaccine candidates	1.9835	tgev infection	1.9878
mass spectrometry	1.9138	host species	1.9833	cs dna	1.9878
national natural	1.9138	respiratory viral infections	1.9832	risk perception	1.9875
rift valley	1.9138	endothelial cells	1.9829	s1 protein	1.9875
science china	1.9138	sequence data	1.9829	ring vaccination	1.9875
valley fever	1.9138	dna viruses	1.9826	syrian hamster	1.9873
respiratory virus infections	1.913	host innate	1.9826	wild mice	1.9873
syndrome coronavirus	1.9096	parainfluenza virus	1.9824	yellow fever	1.9873
classical swine fever virus	1.9087	tract infections	1.9822	climate change	1.9873
b cells	1.9074	south korea	1.9821	public health services	1.9873
host response	1.9074	acute respiratory infection	1.9817	index patients	1.9872
science foundation of china	1.9074	reproduction number	1.9816	small rna	1.9872
viral proteins	1.9074	surveillance system	1.9816	ic activity	1.9871
virus disease	1.9065	causative agent	1.9813	ebola virus disease	1.9868
clinical infectious diseases	1.9048	multiple sclerosis	1.9811	rna chaperone	1.9867
world health organization	1.9048	rsv infection	1.9809	caco-2 cells	1.9867
antiviral agents	1.9001	cellular proteins	1.9808	m2 channel	1.9865
cell culture	1.9001	west nile virus	1.9806	overlapping genes	1.9865
pulmonary disease	1.9001	respiratory diseases	1.9805	nasal mucosa	1.9865
study protocol	1.9001	tgev infection	1.9805	hepatitis e	1.9865
dengue virus	1.8946	e protein	1.9802	genetic drift	1.9865
public health	1.893	gene expression	1.9801	a7 gfp	1.9865
rna replication	1.8915	structural proteins	1.9799	tumor cells	1.9864
japanese encephalitis	1.8902	acute respiratory tract	1.9792	tanguticum nanoparticles	1.9864
syndrome coronavirus infection	1.8864	hand hygiene	1.9792	cfu ml	1.9864
human respiratory syncytial virus	1.8841	disease transmission	1.9788	ward closure	1.9861
synonymous codon usage	1.8824	human rhinovirus	1.9785	case definitions	1.9861
clinical infectious	1.8813	bacterial infections	1.9781	richards model	1.9861
health organization	1.8813	cancer cells	1.9781	epimedium koreanum	1.9861
severe pneumonia	1.8813	dna vaccines	1.9777	ms2 plp	1.986
dengue virus infection	1.8772	type iii	1.9777	gene therapy	1.9859
clinical samples	1.8768	viral pathogenesis	1.9773	integrin b3	1.9859
classical swine fever	1.8744	zoonotic diseases	1.9773	cardiovascular diseases	1.9859
human antibody	1.869	early detection	1.9765	fourth site	1.9859
lassa virus	1.869	lung cancer	1.9756	serial interval	1.9858
pilot study	1.869	nile virus	1.9756	trm cells	1.9858
avian influenza viruses	1.8667	human disease	1.9751	electronic supplementary material	1.9857
human respiratory syncytial	1.8667	rnase l	1.9751	emergency nurses	1.9856
international health regulations	1.8667	health systems	1.9746	pet substrate	1.9856
hepatitis c virus infection	1.8661	incubation period	1.9746	fcov type	1.9856
infectious bronchitis virus strain	1.8661	rabies virus	1.9746	s1 text	1.9856
vaccine development	1.8601	adaptive immunity	1.9741	global health research	1.9854
protects hepatocytes from type i	1.8564	multiplex pcr	1.9741	ace2 activity	1.9853
type i interferon signaling disrupts	1.8564	nk cells	1.9741	β 6 ko	1.9853
adaptive immunity	1.8538	feline coronavirus	1.9735	global health	1.9852
adenovirus type	1.8538	human populations	1.9735	ham tsp	1.9851
nonhuman primates	1.8538	common cold	1.9723	blood culture	1.9849

Table 7 Best ranked terms extracted from Paper1 using F-TFIDF-C_M

C-Value					
Title Corpus		Abstract Corpus		Content Corpus	
terms	rank	terms	rank	terms	rank
respiratory syndrome	386.7309	public health	1393.182	t cells	2063.1457
virus infection	366.1263	respiratory syndrome	1095.2091	public health	1644.7156
porcine epidemic diarrhea virus	329.7138	infectious diseases	952.5625	amino acid	1409.82415
porcine epidemic diarrhea	318.0	immune response	908.1835	immune response	1400.94835
epidemic diarrhea virus	306.0	immune responses	841.6151	influenza virus	1185.8689
east respiratory syndrome	284.0	influenza virus	841.6151	immune responses	1056.536
middle east	261.5188	t cells	803.576	t cell	1056.37753
epidemic diarrhea	256.7639	virus infection	760.7811	gene expression	1050.6716
diarrhea virus	245.6692	respiratory tract	727.4978	viral replication	1021.5083
infectious diseases	245.6692	viral infection	668.8542	infected cells	939.72426
respiratory syndrome coronavirus	240.0	viral replication	665.6843	cell lines	897.4057
influenza a	225.0647	viral infections	640.3249	viral infection	888.6884
public health	209.2151	east respiratory syndrome	638.0	virus infection	872.68035
syndrome coronavirus	191.7805	respiratory syndrome coronavirus	636.0	amino acids	866.816
porcine epidemic	190.1955	middle east	630.8151	mg ml	824.4975
influenza virus	182.2707	gene expression	627.6452	infectious diseases	822.27855
respiratory tract	180.6857	infectious disease	613.3805	present study	812.45177
middle east respiratory syndrome	174.1446	rna viruses	603.8707	respiratory tract	812.13477
middle east respiratory	170.0	present study	575.3414	epithelial cells	759.03855
respiratory syncytial virus	166.0	respiratory viruses	551.567	previous studies	732.41119
infectious bronchitis	160.0812	acute respiratory syndrome	516.0	room temperature	714.3426
infectious disease	156.9113	t cell	513.5279	cell culture	673.60907
infectious bronchitis virus	156.0	syndrome coronavirus	511.9429	additional file	657.75946
east respiratory	136.3068	porcine epidemic diarrhea	506.0	viral infections	635.72848
syncytial virus	134.7218	95percent ci	502.4331	immune system	617.97689
avian influenza	131.5519	viral rna	499.2632	respiratory syndrome	617.3429
respiratory viruses	131.5519	amino acid	489.7534	cell line	611.16155
east respiratory syndrome coronavirus	130.028	respiratory syncytial virus	472.0	infectious disease	607.04063
middle east respiratory syndrome coronavirus	129.2481	cell lines	443.7895	μ g ml	576.13388
influenza a virus	126.0	respiratory infections	426.3549	western blot	568.36754
bronchitis virus	125.212	epithelial cells	424.77	rnase l	565.0391
respiratory infections	125.212	virus replication	420.0151	virus replication	560.6012
systematic review	125.212	polymerase chain reaction	408.0	cell surface	543.9591
ebola virus	120.4572	epidemic diarrhea virus	406.0	xx	542.0572
acute respiratory	117.2872	epidemic diarrhea	402.5805	host cell	539.83825
viral infections	117.2872	host cell	396.2406	codon usage	523.03765
virus replication	115.7023	syncytial virus	378.806	viral proteins	520.6601
open access	109.3624	porcine epidemic diarrhea virus	376.1524	respiratory viruses	515.4298
zika virus	109.3624	antiviral activity	374.0512	nk cells	503.2256
respiratory tract infections	102.0	risk factors	374.0512	time points	497.8367
viral infection	101.4376	immune system	369.2963	influenza viruses	492.7648
immune response	99.8526	ebola virus	364.5414	important role	491.0213
hepatitis c virus	98.0	chain reaction	355.0316	allergic rhinitis	486.5835
gene expression	96.6827	influenza viruses	348.6918	antiviral activity	481.3531
pandemic influenza	96.6827	infected cells	347.1068	global health	473.9038
respiratory syndrome virus	96.0	diarrhea virus	340.7669	mg kg	470.0998
epithelial cells	95.0978	host cells	334.4271	frequency domain	469.1489
complete genome	93.5128	important role	331.2572	control group	466.13749
syndrome virus	93.5128	phylogenetic analysis	331.2572	viral load	465.34499
virology journal	93.5128	polymerase chain	331.2572	binding site	459.6391
hepatitis c	91.9278	respiratory disease	326.5023	expression levels	453.6162
immune responses	90.3429	avian influenza	324.9173	hong kong	450.7237
genome sequence	88.7579	respiratory tract infections	320.0	clinical signs	448.8613
dengue virus	87.1729	infectious bronchitis	285.2933	protein expression	448.2274
molecular sciences	84.0029	cell culture	272.6136	wild type	446.7833
type i	84.0029	hepatitis c virus	268.0	endothelial cells	441.4120
acute respiratory syndrome	84.0	health care	264.6887	table s1	438.4006
complete genome sequence	84.0	zika virus	264.6887	flow cytometry	437.4496
human coronavirus	82.4181	infectious bronchitis virus	260.0	saudi arabia	433.4872
respiratory infection	82.4181	tract infections	258.3489	viral genome	433.3992
case report	80.8331	hepatitis c	255.179	negative control	433.2230
tract infections	80.8331	innate immune response	252.0	–	431.7890
risk factors	79.2481	monoclonal antibodies	248.8391	cell types	431.1098
spike protein	77.6632	viral genome	247.2542	viral entry	427.9399
t cell	77.6632	type i	242.4993	cell death	425.24544
acute respiratory infections	76.0	central nervous system	242.0	er stress	423.185
coronavirus infection	74.4932	amino acids	239.3293	significant differences	420.6490
rna viruses	74.4932	animal models	237.7444	health care	420.4905
severe acute respiratory	72.0	real-time pcr	236.1594	tcid 50	417.3734
sars coronavirus	71.3233	dengue virus	232.9895	cathepsin l	410.5053
isothermal amplification	69.7384	viral load	232.9895	risk factors	408.9203
respiratory disease	69.7384	world health organization	232.0	positive selection	405.7504
bmc public health	66.0	cell line	231.4045	cell cycle	400.9955
disease virus	64.9835	viral proteins	229.8196	nucleotide sequences	397.8256
t cells	63.3985	nervous system	226.6496	plasma membrane	393.5990
influenza viruses	61.8135	wide range	223.4797	intensive care	392.2782
acute respiratory infection	60.0	virus infections	221.8948	host cells	384.82889
type i interferon	60.0	middle east respiratory syndrome	220.5832	hand hygiene	383.5609
journal frontiers	58.6436	immunodeficiency virus	218.7248	significant difference	382.6099
fever virus	57.0587	spike protein	218.7248	immune cells	381.02498
respiratory syncytial	57.0587	life cycle	217.1399	reference genes	380.3909
severe acute	57.0587	recent years	217.1399	hiv aids	377.2211
respiratory tract infection	56.0	codon usage	215.5549	avian influenza	376.8688
antiviral activity	55.4737	viral pathogens	215.5549	serum samples	375.8625
bmc infectious	55.4737	pandemic influenza	213.9699	body weight	375.0021
hong kong	55.4737	clinical signs	212.385	figure 1a	374.0511
viral replication	55.4737	dendritic cells	209.2151	membrane fusion	374.0511
virus infections	55.4737	acute respiratory syndrome coronavirus	208.9735	clinical trials	373.8750
bmc infectious diseases	54.0	bronchitis virus	207.6301	time point	373.3719
respiratory viral infections	54.0	endoplasmic reticulum	207.6301	protein synthesis	369.2962
case study	53.8887	rna virus	207.6301	dengue virus	367.7113
dendritic cells	53.8887	saudi arabia	207.6301	e protein	367.7113
mini review	53.8887	innate immunity	206.0451	high levels	365.3339
rna virus	53.8887	recent studies	206.0451	virus particles	364.5414
transmissible gastroenteritis	53.8887	economic losses	204.4602	target cells	362.5601
bmc public	52.3038	porcine epidemic	204.4602	viral particles	360.4204
monoclonal antibodies	52.3038	world health	204.4602	dendritic cells	357.5675
creative commons cc-by 4	51.0824	global health	202.8752	total number	356.4580
influenza pandemic	50.7188	type 1	202.8752	cancer cells	356.0883
type 1	50.7188	vaccine development	201.2902	disease control	355.2957

Table 8 Best ranked terms extracted from Paper1 using C-Value

sous_corpus	measure	term	domain relevant	COVID-19 surveillance	syndromic surveillance	Incomplet disease name
title	C-value	respiratory syndrome coronavirus	n	n	n	y
		porcine epidemic diarrhea	y	n	n	n
		syndrome coronavirus	n	n	n	y
		epidemic diarrhea virus	n	n	n	y
		acute respiratory syndrome	n	n	n	y
		public access	n	n	n	n
		diarrhea virus	n	n	n	y
		infectious bronchitis	y	n	y	n
		acute respiratory	n	n	n	y
		bronchitis virus	y	n	y	n
	F-TFIDF-C	journal pre-proof	n	n	n	n
		virology journal	n	n	n	n
		influenza pandemic	y	n	n	n
		coronavirus spike	y	y	n	n
		bmc public health	n	n	n	n
		influenza virus infection	y	n	n	n
		emerging infectious	y	n	y	n
		porcine circovirus type	n	n	n	y
		codon usage	n	n	n	n
		respiratory syndrome	n	n	n	y
abstract	C-value	acute respiratory syndrome	n	n	n	y
		respiratory syndrome coronavirus	n	n	n	y
		east respiratory syndrome	n	n	n	y
		syndrome coronavirus	n	n	n	y
		present study	n	n	n	n
		chain reaction	n	n	n	n
		syncytial virus	n	n	n	y
		porcine epidemic diarrhea	y	n	n	n
		polymerase chain	n	n	n	n
	F-TFIDF-C	virus infections	y	n	y	n
		porcine epidemic	n	n	n	y
		clinical samples	n	n	n	n
		codon usage	n	n	n	n
		mers-cov infection	y	y	n	n
		pandemic influenza	y	n	n	n
		viral entry	y	n	y	n
		95percent confidence interval	n	n	n	n
		immune cells	n	n	n	n
		influenza pandemic	y	n	n	n
		sono stati	n	n	n	n
content	C-value	infected cells	n	n	n	n
		respiratory syndrome	n	n	n	y
		present study	n	n	n	n
		individual components	n	n	n	n
		essential medicines	n	n	n	n
		previous studies	n	n	n	n
		de los	n	n	n	n
		functional task	n	n	n	n
		der schwangerschaft	n	n	n	n
	F-TFIDF-C	health emergency	y	n	y	n
		membrane rafts	n	n	n	n
		pcr products	n	n	n	n
		afa dr	n	n	n	n
		cod trypsin	n	n	n	n
		2c atpase	n	n	n	n
		naked mole	n	n	n	n
		intracellular delivery	n	n	n	n
		close contact	y	n	y	n
		final dataset	n	n	n	n
		respiratory syndrome	n	n	n	y
title + abstract	C-value	acute respiratory syndrome	n	n	n	y
		respiratory syndrome coronavirus	n	n	n	y
		east respiratory syndrome	n	n	n	y
		syndrome coronavirus	n	n	n	y
		syncytial virus	n	n	n	y
		porcine epidemic diarrhea	y	n	n	n
		antiviral activity	y	n	n	n
		acute respiratory syndrome coronavirus	n	n	n	y
		infectious bronchitis	y	n	n	n

Table 9 Expanded terms from Table 6. Each term has been evaluated by an expert according 4 criteria: domain relevant, COVID-19 surveillance, syndromic surveillance, incomplet disease name (y: yes, n: no)

influenza virus	**evaluation**	**respiratory infections**	**evaluation**	**infectious diseases**	**evaluation**
influenza a/wsn/33 virus	not relevant	respiratory virus infections	relevant	diseases relates to infectious	relevant
viruses and conventional influenza	relevant	respiratory viral infection	relevant	disease called feline infectious	relevant
virus remains the influenza	lack of context	infection by respiratory	relevant	infectious animal diseases	relevant
influenza by virus	not relevant	infections of the respiratory	relevant	infectious enteric diseases	relevant
influenza vaccine virus	relevant	infect respiratory	relevant	disease without being infectious	not relevant
virus and canine influenza	relevant	infections are respiratory	relevant	disease has an infectious	lack of context
virus influenza	relevant	infected with respiratory	relevant	infectious disease	relevant
viruses such as influenza	relevant	infection with other respiratory	relevant	disease named it infectious	lack of context
influenza b viruses	relevant	respiratory virus infection	relevant	infectious swine diseases	relevant
viruses and emerging influenza	relevant	infection transmitted via respiratory	relevant	disease models for infectious	lack of context
viral infections	**evaluation**	**sars coronavirus**	**evaluation**	**incubation period**	**evaluation**
viral bronchopulmonary infection	relevant	coronavirus is urbani sars	not relevant	incubating period	relevant
virally infected	relevant	coronavirus of 18 sars	not relevant	periods of incubation	relevant
viral respiratory infections	relevant	coronavirus that causes sars	relevant	incubation periods	relevant
infection and encounter virally	lack of context	coronavirus named sars	relevant	period of incubation	relevant
infection or viral	lack of context	coronavirus related to sars	relevant	period than incubation	lack of context
viral skin infection	relevant	coronavirus isolated from sars	relevant	period and incubation	lack of context
virals infection	relevant	coronavirus responsable du sars	relevant	incubation for period	lack of context
infection with one viral	relevant	sars -associated coronavirus	relevant	period and incubating	lack of context
viral opportunistic infections	relevant	sars human coronavirus	relevant	period covering an incubation	relevant
infection at high viral	lack of context	sars and coronavirus	relevant	period of extrinsic incubation	relevant

Table 10 60 terms randomly selected from FASTR variants (Section 4.2)

B IMAGES

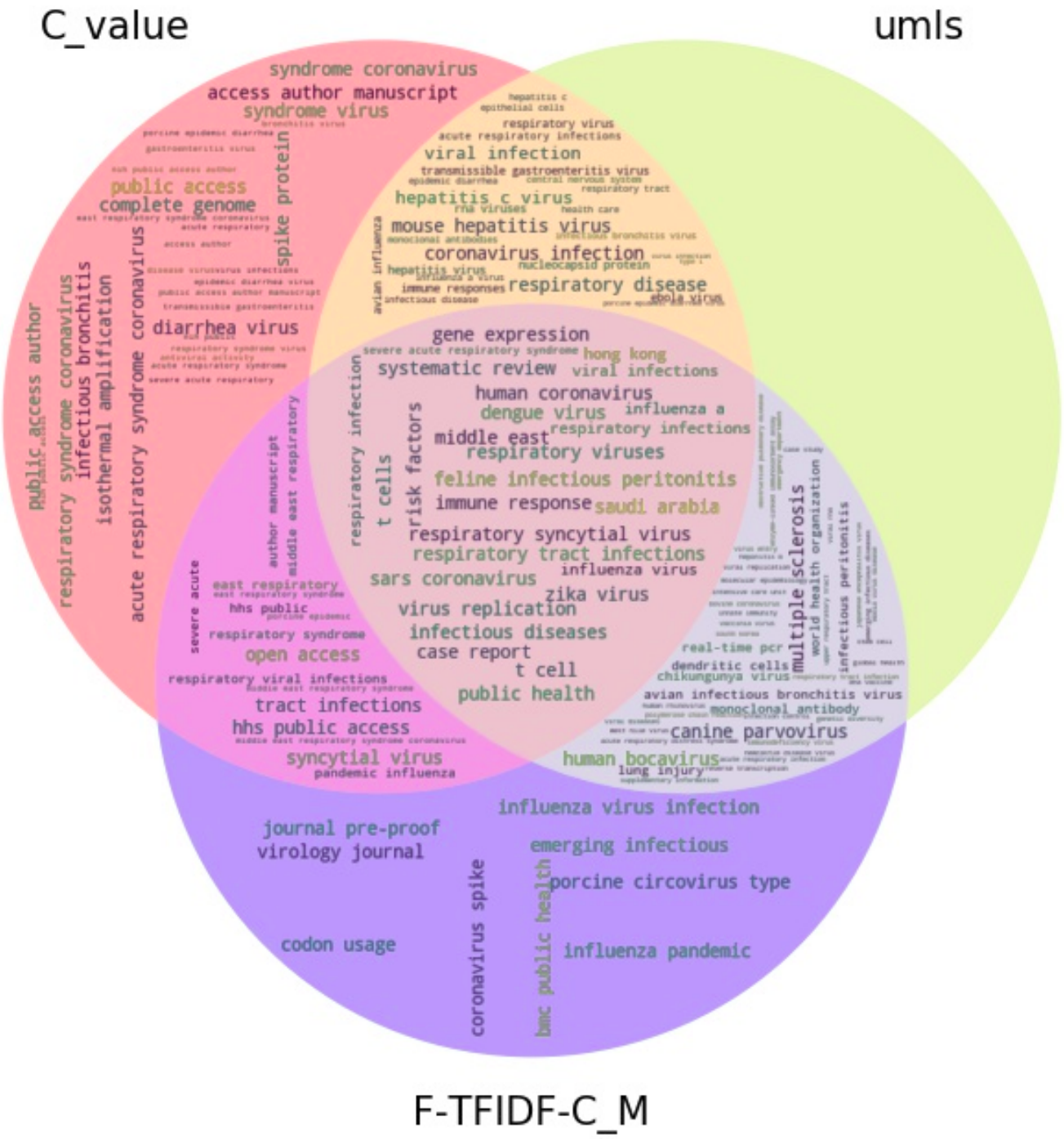

Fig. 10 Distribution of concepts according to the measures and their presence in the UMLS Metathesaurus: from Papers2-title corpus

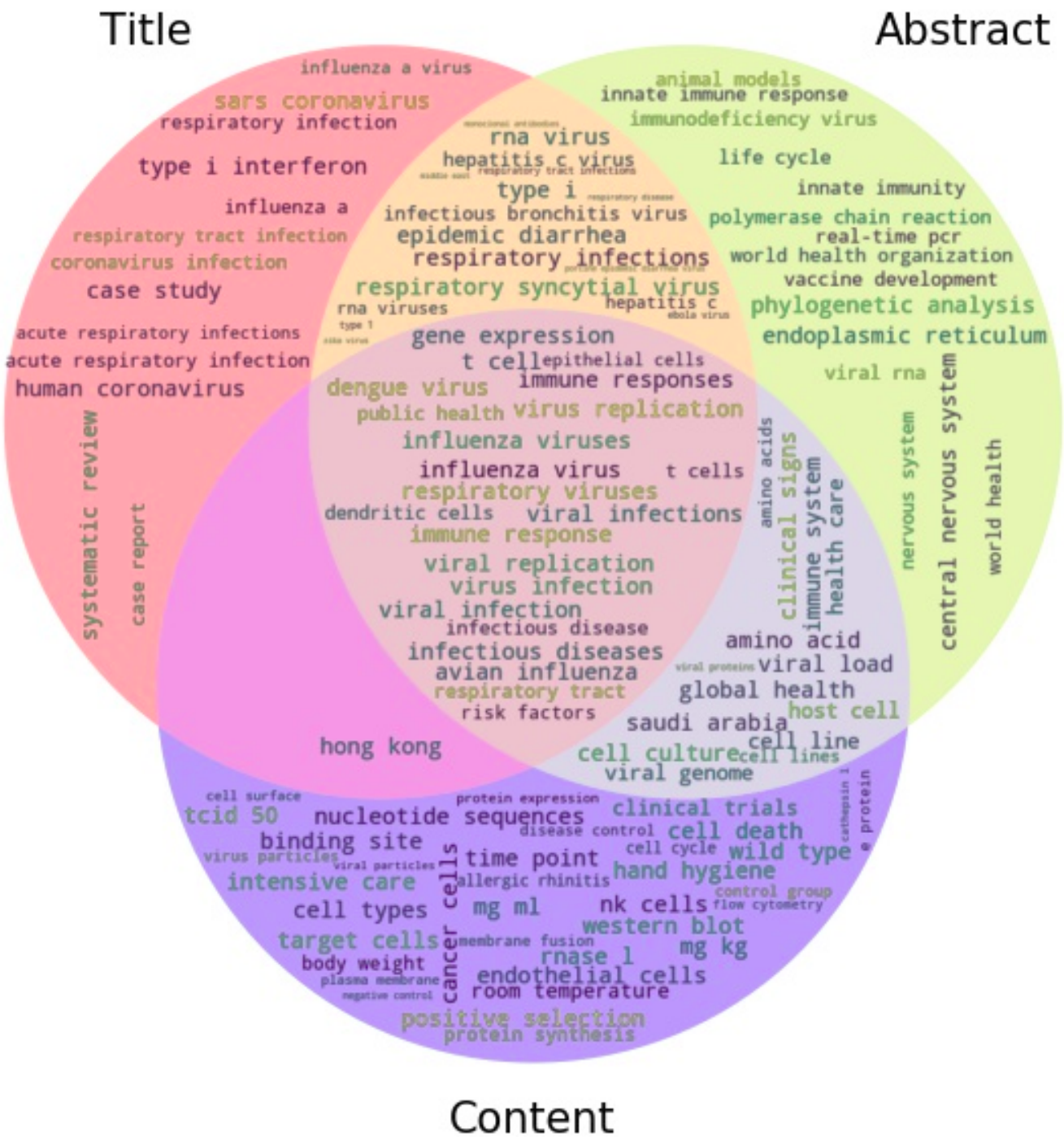

Fig. 11 Distribution of representative concepts when taking multiple corpora into account using C_Value: Papers1 corpora