Scientometrics (2021) 126:7445–7475
https://doi.org/10.1007/s11192-021-04085-9

An approach for detecting the commonality and specialty between scientific publications and patents

Shuo Xu[1] · Ling Li[1] · Xin An[2] · Liyuan Hao[1] · Guancan Yang[3]

Received: 15 August 2020 / Accepted: 21 June 2021 / Published online: 5 July 2021

Abstract
Scientific publications and patents are usually viewed as respective proxies of scientific research and technical development. There is considerable effort spent towards establishing topic linkages between science and technology with the lexical- or topic-based approaches. However, due to the heterogeneity between scholarly articles and patents in terms of purpose, statement, and quality, the performance is not satisfactory. To understand the difficulties of topic linkages and improve the performance, a framework is proposed to detect the commonality and specialty between scientific publications and patents from the two perspectives: linguistic characteristics and thematic structures. Extensive experimental results on the DrugBank dataset discover five commonness and five significant differences in terms of linguistic characteristics. For example, nouns are used most frequently among them, and scientific publications contain more word tokens than patent documents, but patents have usually longer sentences and use more clauses. In the meanwhile, common and special thematic structures are also uncovered between scientific publications and patents. The themes about general description in the pharmaceutical field are shared by two heterogeneous resources. The scientific publications tend to explain the disease mechanism and the medication content, while patents bias towards the preparation and practical application of drugs.

Keywords Linguistic characteristics · Stopword identification · Multi-collection topic model · Scientific publication · Patent

Introduction

Throughout the development of modern science and technology, scientific findings and technological innovations have been showing a novel pattern. Science and technology, as two forces that influence the direction of Scientific & Technological (S&T) development, interpenetrate and interweave. Similar to a double stranded DNA, they have been forming a spiral upward developmental trend (Brooks, 1994; Xu et al., 2020). It has been shown that the intersection and integration of science and technology is often an

✉ Xin An
anxin@bjfu.edu.cn

Extended author information available on the last page of the article

opportunity for major technological innovations and has become an important driving force for technologies to emerge (Albert, 2016; Andy, 2007; Lee et al., 2011). Therefore, it is increasingly important to explore and exploit the linkages between science and technology.

In the literature, scientific publications and patent documents are usually viewed as respective proxies of scientific research and technical development (Calero-Medina & Noyons, 2008; Dubaric et al., 2011; Xu et al., 2019a). Last 2 decades witnessed significant progress in the field of the science-technology linkage analysis ever since Narin et al. (1997). Several perspectives have been investigated, such as mutual citations between scientific publications and patents (Gao et al., 2012a; Huang et al., 2015; Verbeek et al., 2002), and academic inventors (researchers with authorship and inventorship roles) (Forti et al., 2007; Wang & Guan, 2011). As a matter of fact, limited number of mutual citations and academic inventors prevents many scholarly articles and patents from being included in the analyzed dataset (Shibata et al., 2011; Takano et al., 2016).

With the rapid development of machine learning technique and the rise of Open Access (OA) movement, it is more convenient to access the textual content of scholarly articles and patent documents. Therefore, there is considerable effort spent towards establishing the linkages between science and technology with the lexical- or topic-based approaches (Bassecouolard & Zitt, 2004; Shibata et al., 2011; Xu et al., 2019a). On closer examination, one can see that previous studies basically adopted a similar non-joint framework as follows. After thematic structures are extracted respectively from scientific publications and patents, the similarities between themes from different resources are calculated, and then topic linkages are constructed. However, due to the heterogeneity between scholarly articles and patents in terms of purpose, statement, and quality (Xu et al., 2019a), the following challenges will be faced.

As we all know, different types of documents orient different population and carry different purposes. One can communicate scientific finding to the relevant research community and general public through a scholarly article. According to requirements of patent laws, a patent document should disclose the details of an invention as much as possible, but also protect the resulting invention from infringement (An et al., 2021; Chen et al., 2020). In this way, different language specifications and writing styles are followed by scientific publications and patents, which make these two resources to show different linguistic characteristics.

In addition, a common manually curated stopword list is not applicable in this situation, since it is very possible that a term can be viewed as a stopword in the academic articles, but not as a stopword in the patent documents, and vice versa (Gerlach et al., 2019). In other words, these context-dependent terms are ambiguous. Last but not least, as mentioned in Shibata et al., (2010, 2011) and Xu et al. (2012), the performance of conventional non-joint framework is not satisfactory. In our opinion, main reason should attribute to non-comparable thematic structures with different distributions from scientific publications and patents.

To meet these challenges, understand the difficulties of topic linkages, and improve the performance of topic linkages, this study devotes to developing a framework for detecting the commonality and specialty between scientific publications and patents. In more details, the HMM-LDA (Hidden Markov Model-Latent Dirichlet Allocation) model (Griffiths et al., 2004) is introduced here to automatically recognize the stopwords from scholarly articles and patent documents. Then, linguistic characteristics are measured with several syntactic and lexical complexity indicators. In the end, a revised joint topic model for multi-corpora, Common and Distinctive Topic Model (CDTM) (Hua et al., 2020), is used

to extract the common and special topics among scientific publications and patents. The following summarizes main contributions of this work:

- An approach is proposed to detect the commonality and specialty between scientific publications and patents from the perspectives of linguistic characteristics and thematic structures.
- To deal with ambiguous terms in different contexts, an advanced stopword identification method, HMM-LDA model, is adopted to separate informative words from non-informative ones.
- A joint topic model for multi-corpora, CDTM model, is revised to simultaneously discover the shared thematic structures among articles and patents as well as different number of unique themes specific to each resource.

The rest of the article is arranged as follows. After related work is briefly reviewed, our research framework for detecting the commonality and specialty between scientific publications and patents are described in more details. In our framework, three important modules are involved: stopword identification, syntactic and lexical complexity indicators and a joint thematic structure discovery model. Finally, extensive experimental results on the DrugBank dataset uncover several significant differences in terms of linguistic characteristics, and common and special thematic structures between scientific publications and patents.

Literature review

Before delving into more specifies, discussions of the literature pertinent to stopword identification methods, indicators for measuring linguistic characteristics and topic models for multiple corpora are in order.

Stopword identification methods

The stopwords are common words which would appear to be of little value for an interested task, but they often account for a large part of textual data. For purpose of improving the efficiency and accuracy of text analysis, the removal of stopwords serves as one of important preprocessing steps.

A popular method for filtering stopwords resorts to a pre-built stopword list. Though, there is no single universal list of stopwords used by all natural language processing (NLP) tools. To accommodate different applications, several approaches for constructing a stopword list have been raised, such as heuristic methods based on the number of occurrences (most and least frequent words), document frequency, and term frequency and inverse document frequency (TF-IDF) (Christopher, 1989; Gerard, 1963; Salton & Yang, 1973) and those built on the information theory (Gerlach et al., 2019; Montemurro & Zanette, 2010). As a matter of fact, it is well known that many stopwords are domain-dependent. For instance, the term *DNA* may be a stopword in the domain of *biopharmacy*, but a keyword in *biomolecular computing*. For purpose of extracting these domain-specific stopwords, several dedicated methods are also proposed in the literature (Makrehchi & Kamel, 2008, 2017; Seki & Mostafa, 2005).

On closer examination, it is not difficult to see that the above methods do not still deal with ambiguous terms. By ambiguity, we mean that a term is non-informative in some contexts, but informative in other contexts, such as *He* in the following excerptions "…, but the mechanism of *He* inserting into C60 cage at explosive conditions was not clear…" and "…*He* did not have electrocardiogram (ECG) changes suggestive of myocardial pathology, …". The HMM-LDA developed by Griffiths et al. (2004) tries to introduce syntactic structure information into a topic model by the Hidden Markov Model (HMM) (Rabiner, 1989) and the LDA (Latent Dirichlet Allocation) (Blei et al., 2003). This model can effectively distinguish between content and function word tokens. Thus, it is able to recognize the domain-specific stopwords, but also tell an informative meaning from a non-informative one for an ambiguous term. Hence, the HMM-LDA model is utilized here to identify and filter stopwords and background function words in scientific publications and patents.

Indicators for measuring linguistic characteristics

To measure quantitatively linguistic characteristics of textual data, many indicators have been developed in the literature from the perspectives of complexity, accuracy, and fluency (Brants, 2000; Brown et al., 1993; Lu et al., 2019). Since it is very difficult to capture linguistic characteristics of academic articles and patent documents from the perspectives of accuracy and fluency (Lu et al., 2019), the perspective of linguistic complexity is preferred in this study. The indicators for linguistic complexity can be further grouped into two categories: (a) syntactic complexity (Ferris, 1994; Lu, 2010; Ortega, 2003) and (b) lexical complexity (Ellis & Yuan, 2004; Kormos, 2011; Lu et al., 2019). The syntactic complexity focuses on the difference of language expression at the sentence level, while the lexical complexity is related to the measurement from the perspective of words.

For ease understanding, several indicators are illustrated in Table 1, such as *length*, *sentence length* and *sentence complexity* for syntactic complexity, and *lexical diversity*, *lexical density*, and *lexical sophistication* for lexical complexity. From Table 1, it is easy to see that each document is considered as a whole when calculating the resulting indicator. That is, structure of each document (such as *title*, *abstract*, *body* and so on) is discarded directly. As mentioned in "Introduction" section, this article aims to exploit comprehensively the commonalities and specialties between scientific publications and patents. Hence, these indicators are extended to measure linguistic characteristics of *title* and *abstract* parts of scholarly articles and patent documents (cf. Table 2).

Topic models for multi-corpora

The topic models are a suite of algorithms for discovering the hidden thematic structures from a large collection of documents (Blei, 2012). Ever since the standard LDA model (Blei et al., 2003), many topic models have put forward in the literature (Chen et al., 2015; Paul & Girju, 2010; Zhai et al., 2004). Furthermore, a large number of successful applications have shown the power of these models in various text corpora (An et al., 2014; Wang et al., 2018; Xu et al., 2019b). For more elaborate and detailed surveys, we refer the readers to Blei (2012) and Zhang et al. (2014).

However, on closer examination, one can see that the majority of models are unable to effectively discover the commonality and specialty amongst multiple corpora. A naïve solution is to run a conventional topic model separately on different corpora, and then followed by additional post-processing techniques such as topic pair mapping (Xu

Table 1 Several indicators for measuring syntactic and lexical complexities of a given corpus

Category	Indicator	Description	Formula
Syntactic complexity	Length	Avg. number of word tokens in each document	$L = \frac{1}{M}\sum_{m}^{M} N_m$
	Sentence length	Avg. number of word tokens in each sentence of a document	$SL = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{S_m}\sum_{s=1}^{S_m} N_{m,s}\right)$
	Sentence complexity	Avg. number of clauses in each sentence of a document	$SC = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{S_m}\sum_{s=1}^{S_m} C_{m,s}\right)$
Lexical complexity	Lexical diversity	Avg. ratio of unique words of a document	$DIV = \frac{1}{M}\sum_{m=1}^{M}\frac{V_m}{N_m}$
	Lexical density	Ratio of word tokens with the category c of each document	$DEN = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N_m}\sum_{n=1}^{N_m}\delta\left(w_{m,n}\in c\right)\right)$
	Lexical sophistication	Avg. length of word tokens with the category c of each document	$SOP = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N_m}\sum_{n=1}^{N_m}\delta\left(w_{m,n}\in c\right)\left\|w_{m,n}\right\|\right)$

M and V denote respectively the number of documents and unique words in a corpus; N_m, V_m and S_m are the number of word tokens, unique words, and sentences in the document m, respectively; $N_{m,s}$ and $C_{m,s}$ represents respectively the number of word tokens and clauses in the sentence s of the document m; $w_{m,n}$ denotes the nth word token in the document m, and $|w_{m,n}|$ means the length of word token $w_{m,n}$; The Dirac delta function $\delta(x) = 1$ if the condition x holds, 0 otherwise.

Table 2 Indicators for measuring syntactic and lexical complexity of scientific publications and patents

Indicator	Description
TL	Avg. number of word tokens in each title
AL	Avg. number of word tokens of each abstract
ASL	Avg. number of word tokens in each sentence of an abstract
TSC	Avg. number of clauses in each sentence of a title
ASC	Avg. number of clauses in each sentence of an abstract
TDIV	Avg. ratio of unique words of a title
ADIV	Avg. ratio of unique words of an abstract
TDEN	Avg. ratio of word tokens with the category c of each title
ADEN	Avg. ratio of word tokens with the category c of each abstract
TSOP	Avg. length of word tokens with the category c of each title
ASOP	Avg. length of word tokens with the category c of each abstract

et al., 2012, 2019a). In fact, the performance of such non-joint method is inadequate (Shibata et al., 2010, 2011; Xu et al., 2012), since non-comparable topics with different distributions are generated from different corpora. This enables it difficult to align the discovered thematic structures from different corpora by similarity calculation (Xu et al., 2019a). More specifically, many topic pairs with top similarity values are not taken to be semantically similar to each other by domain experts.

Another intuitive solution is to combine these corpora to a greater corpus in the first place, and then to uncover thematic structures on this combined corpus with a traditional topic model (such as LDA). Though the same topic distributions can be obtained for multiple corpora, this solution implicitly assumes that the documents across multiple corpora are exchangeable (Wang et al., 2009; Xu et al., 2012). As a matter of fact, this assumption is not appropriate for our situation, since this work wants to respect the boundaries of the corpora but accounting for and estimating their commonality and specialty.

The ccMix (cross-collection Mixture) model (Zhai et al., 2004) as a first step in this direction tries to perform cross-corpus clustering and within-corpus clustering simultaneously to discover the latent common themes across corpora within the probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) framework. The Markov Topic Model (MTM) (Wang et al., 2009) can capture both the internal topic structures within each corpus and the relations between topics across the corpora. Nonetheless, the ccMix model is not fully Bayesian due to the intrinsic weakness of the pLSA-framework, and the MTM model does not explicitly consider the similarities and differences between corpora. To overcome these limitations, Paul (2009) proposed cross-collection LDA (ccLDA) to capture meaningful word co-occurrence patterns across multiple corpora and model their similarities and differences across multiple corpora.

Then, the ccLDA model is extended to several variants, such as the Topic-Aspect Model (TAM) (Paul & Girju, 2010) for aspect mining from multiple corpora, the scLDA (supervised cross-collection LDA) (Gao et al., 2012b) for the supervised learning scenario, and DTM (Differential Topic Model) (Chen et al., 2015) for the non-parametric version. As the state-of-the-art model in this direction, the CDTM (Common and Distinctive Topic Model) (Hua et al., 2020) is able to discover topics characterizing a particular corpus, as well as maximally exploit the shared information across multiple corpora.

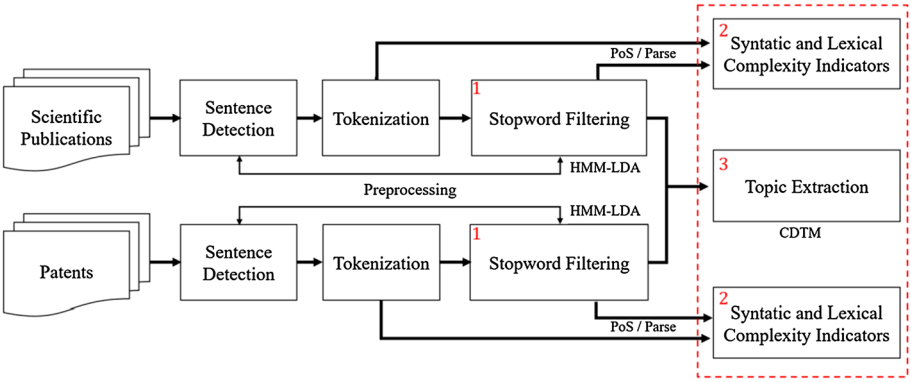

Fig. 1 Research framework for detecting the commonality and specialty between scientific publications and patents

However, the same number of specific topics for each corpus is enforced in Hua et al. (2020). Obviously, this constraint is counter-intuitive, since each corpus should have different number of themes. Hence, this study develops a variant of the CDTM model to loosen this limitation by incorporating a different hyper-parameter for each corpus. Then, this revised CDTM model is utilized to detect the commonality and specialty between scientific publications and patent documents in term of thematic structures.

Research framework and methodology

To detect the commonality and specialty between scientific publications and patents, our research framework mainly consists of three modules, as shown in Fig. 1. After detecting sentence boundaries, and tokenizing each detected sentence, the HMM-LDA model (Griffiths et al., 2004) is used to separate content word tokens from background function ones in scientific publications and patent documents. Then, linguistic characteristics are measured with syntactic and lexical complexity indicators, and thematic structures are simultaneously discovered from academic articles and patents with our revised CDTM model. In the following subsections, these three main modules will be described in more details one by one.

Stopword filtering

Just as everyone's existence has its social meaning, words also play different roles in the expression of text semantics. Griffiths et al. (2004) summarizes the reasons for the appearance of words in sentences into two categories: one is the syntactic function with short-range constraints, making the sentence conform to a certain language specification; the other is to provide the semantic function of long-range constraints to convey the real meaning of an interested sentence. The HMM-LDA model (Griffiths et al., 2004) can take into considerations short-range syntactic dependencies and long-range semantic dependencies between words simultaneously. Thus, ambiguous problem mentioned in "Stopword identification methods" section can be overcome very well when the contexts are considered, such as *all* and *function* in Fig. 3. This is different from the practice of removing stopwords

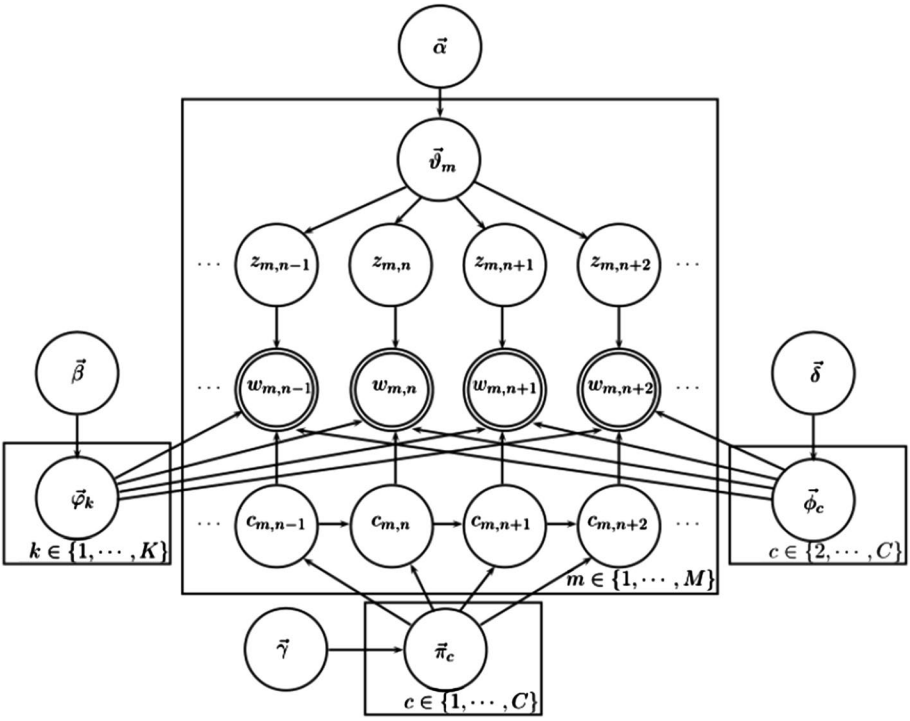

Fig. 2 The graphical model representation of the HMM-LDA model

in many text preprocessing tasks. Here, the HMM-LDA model is used to identify and filter the stopwords and background function words that have no meaning for the topic expression in scientific publications and patents, so that one can obtain content words that contribute to the thematic expression.

More specifically, as the name states, the HMM-LDA model consists of two components: HMM for syntactic function and LDA for semantic function. The graphical model representation for the HMM-LDA model is shown in Fig. 2. The HMM component makes up of the nodes $\vec{c}$, $\vec{w}$, $\{\vec{\pi}_c\}_{c=1}^{C}$ and $\{\vec{\phi}_c\}_{c=2}^{C}$ with the hyper-parameters $\vec{\gamma}$ and $\vec{\delta}$, and the LDA component of the other nodes in Fig. 2. For each word token $w_{m,n}$ in the document m, a topic and class index $z_{m,n}$ and $c_{m,n}$ is associated with it, where $c_{m,n} = 1$ is designated as the semantic class in this study. When $c_{m,n} = 1$ each word token $w_{m,n}$ is drawn randomly from the multinomial distribution $\vec{\varphi}_{z_{m,n}}$, otherwise from the multinomial distribution $\vec{\phi}_{c_{m,n}}$. Each document m is assumed to follow a multinomial distribution $\vec{\vartheta}_m$, and transition probabilities from the class c to other classes follow another multinomial distribution $\vec{\pi}_c$.

For easy understanding the power of the HMM-LDA model, several examples are illustrated in Fig. 3. Syntactic function words and content function words, judged by the HMM-LDA model, are colored in gray and black respectively. For example, it is very obvious that the term *all* in the first excerption does not carry any valuable information, and the term *all* in the second excerption expresses the same meaning as its long form *acute lymphoblastic leukemia*. That is to say, the first mention can be seen as a syntactic function word, and the second as a content function word. Similarly, the term *function* also serves as a different

1	[PMID: 17381384] when **hit** is suspected , prompt cessation of all **heparin** therapy is necessary , along with initiation of **alternative anticoagulant** therapy . [PMID: 25348002] **asparaginase** is a critical agent used to treat acute **lymphoblastic leukemia** (**all**) .
2	[PMID: 17335414] **enrolment** has begun in a Phase I trial evaluating whether **systemically** delivered 131 I - TM - 601 can be used to image metastatic **solid tumors** and primary **gliomas** . [PMID: 7590775] **Antiidiotypic antibodies** bearing the **internal image** of an **antigen** expressed on the surface of the **tumor** seem to be most suited for this purpose .
3	[PMID: 10480573] effect of **thiazinotrienomycin b** , an **ansamycin antibiotic** , on the function of **epidermal growth factor** receptor in human **stomach tumor cells** [PN: US10097388] the first **reference template** may include a **first reference function** , and the **second reference template** may include a second **reference function** in **quadrature** with the **first reference function** .
4	[PN: US6627210] solubility **enhancing components** which aid in solubilizing the alpha - NUMBER - **adrenergic agonist components** . [PMID: 28220701] an **α - aminophenone uptake** inhibitor at plasma **membrane transporters** for **dopamine** (**DAT**) and **norepinephrine** (**NET**) , is a widely prescribed **antidepressant** and **smoking cessation aid** .

Fig. 3 Syntactic function words (in gray color) and content function words (in black color) judged by the HMM-LDA model in several sentences excerpted from scientific publications and patents. The red boxed words serve as a different role depending on the contexts, viz. ambiguous terms

role in the article with PubMed Identifier (PMID) 10480573 and the patent with the patent number (PN) US10097388. Though these terms are ambiguous in term of role expressing text semantics, the HMM-LDA model can deal with them very well according to their contexts.

Similar to Griffiths et al. (2004), the 3rd order HMM is used in this work. The number of topics K and the number of classes C are fixed respectively to 100 and 20, and the symmetric Dirichlet priors α, β, γ, and δ are set at 0.5, 0.01, 0.1, and 0.01, respectively. The Gibbs sampling is run for 1000 iterations, including 200 for the burn-in period.

Syntactic and lexical complexity indicators

The main purpose of this subsection is to measure the commonness and specialties between scientific publications and patents from the perspective of linguistic complexity. The measurement of linguistic complexity is twofold: syntactic complexity and lexical complexity. This work mainly focuses on the linguistic characteristics of *title* and *abstract* parts of scholarly articles and patent documents. Therefore, several variants in Table 1 are utilized here. More specifically, *title length* (*TL*), *abstract length* (*AL*), *sentence length of the abstract* (*ASL*), *title complexity* (*TSC*) and *abstract complexity* (*ASC*) are utilized to measure syntactic complexity, and *lexical diversity* (title: *TDIV*; abstract: *ADIV*), *lexical density* (title: *TDEN*; abstract: *ADEN*) and *lexical sophistication* (title: *TSOP*; abstract: *ASOP*) for lexical complexity measurement. Please check Table 2 for more details. Note that this study utilizes the resulting Part-of-Speech (PoS) to denote the category of each word token. Due to space limitation, only noun, verb, adj., and adv. are considered here.

Topic extraction

Schmiedel et al. (2019) observed the positive correlation between the number of documents and the number of topics. In many real-world applications, the number of documents in each corpus is usually not equal. Therefore, for purpose of discovering the common and special topics between scholarly articles and patent documents, the CDTM

Table 3 Notations used in the original CDTM model and the revised counterpart

Symbol	Description
K_0, K_ℓ	Number of common topics and specific topics for the corpus ℓ
K_1	Number of specific topics for each corpus
M_ℓ	Number of documents in the corpus ℓ
V	Number of unique words in the all documents
L	Number of corpora
$N_{\ell,m}$	Number of word tokens in the document m of the corpus ℓ
$\vec{\theta}_{\ell,m}$	Multinomial distribution of common topics specific to the document m of the corpus ℓ
$\vec{\vartheta}_{\ell,m}$	Multinomial distribution of specific topics specific to the document m of the corpus ℓ
$\vec{\phi}_k$	Multinomial distribution of words for the common topic k
$\vec{\varphi}_{\ell,k}$	Multinomial distribution of words for the specific topic k in the corpus ℓ
$\vec{\lambda}_{\ell,m}$	Bernoulli distribution of preference status for the document m of the corpus ℓ
$z_{\ell,m,n}$	Topic associated with the nth word token in the document m of the corpus ℓ
$x_{\ell,m,n}$	Status associated with the nth word token in the document m of the corpus ℓ
$w_{\ell,m,n}$	n-th word token in the document m of the corpus ℓ
$\vec{\alpha}_0, \vec{\alpha}_1, \vec{\alpha}_\ell, \vec{\beta}_0, \vec{\beta}_1, \vec{\gamma}$	Hyperparameter

model (Hua et al., 2020) is revised in this subsection to deal with the case of different number of special topics. In the original CDTM model and our revised counterpart, several topics which are shared by all corpora are called common topics, while other topics locally owned by each respective corpus are referred to as specific topics.

Table 3 summarizes the notation used in the CDTM model. The graphical model representations of the original CDTM model and the revised version are shown in Fig. 4. By comparing Fig. 4a, b, one can see that our revised model degenerate to the original CDTM model if the following conditions are met. (1) Each corpus has the same number of specific topics (i.e., $K_1 = K_2 = \cdots = K_L$). (2) The hyperparameter is shared by the multinomial distributions of specific topics for each corpus (viz. $\vec{\alpha}_1 = \vec{\alpha}_2 = \cdots = \vec{\alpha}_L$). That is to say, the original CDTM model is a special case of our model. Furthermore, similar to the original model, our model is also able to simultaneously learn both common topics and specific/distinctive topics. In addition, one can describe the revised CDTM model from the viewpoint of generative process as follows.

1. For each common topic $k \in [1, K_0]$, draw $\vec{\phi}_k \sim \mathrm{Dir}\left(\vec{\beta}_0\right)$;
2. For each corpus $\ell \in [1, L]$ and each specific topic $k \in [1, K_\ell]$, draw $\vec{\varphi}_{\ell,k} \sim \mathrm{Dir}\left(\vec{\beta}_1\right)$;
3. For each corpus $\ell \in [1, L]$ and each document $m \in [1, M_\ell]$ in this corpus, draw $\vec{\theta}_{\ell,m} \sim \mathrm{Dir}\left(\vec{\alpha}_0\right)$, $\vec{\vartheta}_{\ell,m} \sim \mathrm{Dir}\left(\vec{\alpha}_\ell\right)$, and $\vec{\lambda}_{\ell,m} \sim \mathrm{Beta}\left(\vec{\gamma}\right)$, respectively;
4. For each word token $n \in [1, N_{\ell,m}]$, in the document m of corpus ℓ,draw $x_{\ell,m,n} \sim \mathrm{Bern}\left(\vec{\lambda}_{\ell,m}\right)$. If $x_{\ell,m,n} = 0$, draw $z_{\ell,m,n} \sim \mathrm{Mult}\left(\vec{\theta}_{\ell,m}\right)$ and then $w_{\ell,m,n} \sim \mathrm{Mult}\left(\vec{\phi}_{z_{\ell,m,n}}\right)$; Otherwise, draw $z_{\ell,m,n} \sim \mathrm{Mult}\left(\vec{\vartheta}_{\ell,m}\right)$ and then $w_{\ell,m,n} \sim \mathrm{Mult}\left(\vec{\varphi}_{\ell,z_{\ell,m,n}}\right)$.

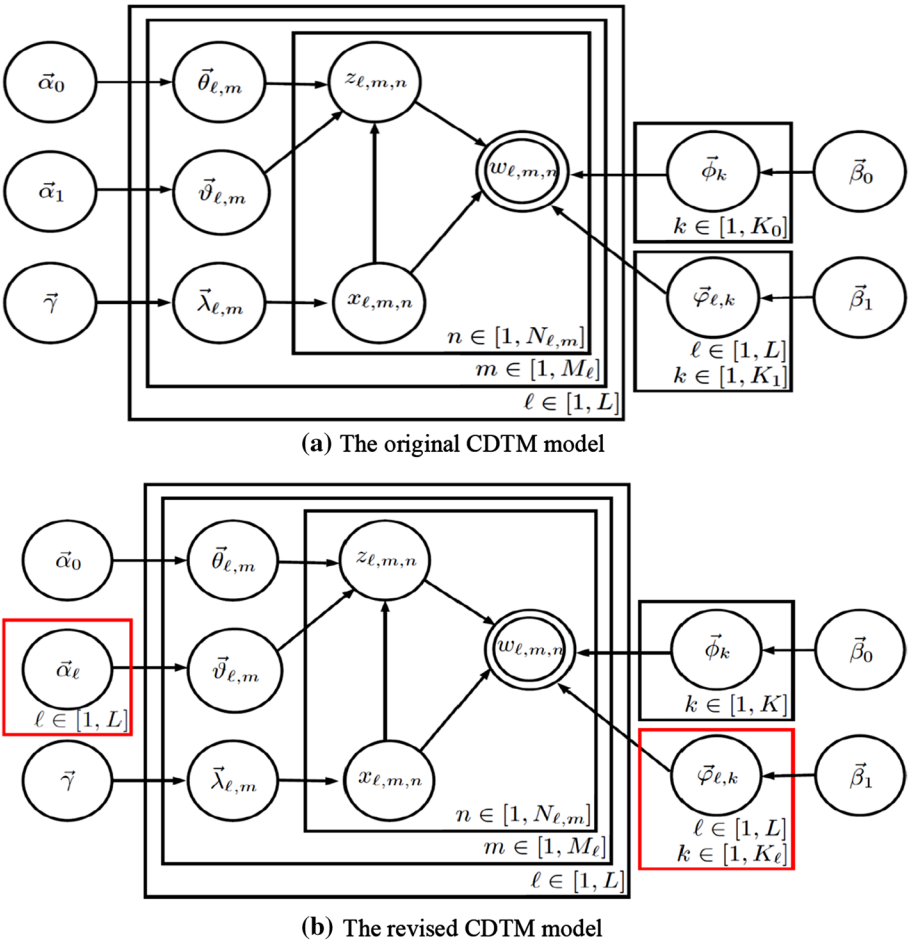

Fig. 4 The graphical model representation of the original CDTM model (**a**) and the revised counterpart (**b**). (Color figure online)

Similar to the original CDTM model (Hua et al., 2020), posterior inference cannot be done exactly in this revised model. The collapsed Gibbs sampling algorithm, a special case of Markov Chain Monte Carlo (MCMC), was originally utilized in Hua et al. (2020) to approximate the posterior of the CDTM model. Hence, the collapsed Gibbs sampling algorithm is also used here. More specifically, in the collapsed Gibbs sampling procedure, two posterior distributions [cf. Eqs. (1), (2)], conditional distributions of the hidden random variables ($z_{\ell,m,n}$ and $x_{\ell,m,n}$) given the observations and other hidden variables, should be calculated. After a derivation, the posterior distribution can be obtained as follows:

$$\Pr\left(z_{\ell,m,n}|\overrightarrow{w},\vec{z}_{\neg(\ell,m,n)},\vec{x},\vec{\alpha}_0,\left\{\vec{\alpha}_\ell\right\}_{\ell=1}^{L},\vec{\beta}_0,\vec{\beta}_1,\vec{\gamma}\right)$$
$$=\begin{cases}\dfrac{\left(n_{0,\ell,m}^{\left(z_{\ell,m,n}\right)}+\alpha_{0,z_{\ell,m,n}}-1\right)}{\sum_{k=1}^{K}\left(n_{0,\ell,m}^{(k)}+\alpha_{0,k}\right)-1}\dfrac{n_{z_{\ell,m,n}}^{\left(w_{\ell,m,n}\right)}+\beta_{0,w_{\ell,m,n}}-1}{\sum_{v=1}^{V}\left(n_{z_{\ell,m,n}}^{(v)}+\beta_{0,v}\right)-1}, & \text{if} \quad x_{\ell,m,n}=0\\ \dfrac{\left(n_{1,\ell,m}^{\left(z_{\ell,m,n}\right)}+\alpha_{1,z_{\ell,m,n}}-1\right)}{\sum_{k=1}^{K}\left(n_{1,\ell,m}^{(k)}+\alpha_{1,k}\right)-1}\dfrac{n_{z_{\ell,m,n}}^{\left(w_{\ell,m,n}\right)}+\beta_{1,w_{\ell,m,n}}-1}{\sum_{v=1}^{V}\left(n_{z_{\ell,m,n}}^{(v)}+\beta_{1,v}\right)-1}, & \text{if} \quad x_{\ell,m,n}=1\end{cases} \tag{1}$$

$$\Pr\left(x_{\ell,m,n}|\overrightarrow{w},\vec{z},\vec{x}_{\neg(\ell,m,n)},\vec{\alpha}_0,\vec{\alpha}_1,\vec{\beta}_0,\vec{\beta}_1,\vec{\gamma}\right)$$
$$=\begin{cases}\dfrac{\left(n_{0,\ell,m}^{\left(z_{\ell,m,n}\right)}+\alpha_{0,z_{\ell,m,n}}-1\right)}{\sum_{k=1}^{K}\left(n_{0,\ell,m}^{(k)}+\alpha_{0,k}\right)-1}\dfrac{n_{\ell,m}^{(0)}+\gamma_0-1}{\sum_{s=0}^{1}\left(n_{\ell,m}^{(s)}+\gamma_s\right)-1}\dfrac{n_{z_{\ell,m,n}}^{\left(w_{\ell,m,n}\right)}+\beta_{0,w_{\ell,m,n}}-1}{\sum_{v=1}^{V}\left(n_{z_{\ell,m,n}}^{(v)}+\beta_{0,v}\right)-1}, & \text{if} \quad x_{\ell,m,n}=0\\ \dfrac{\left(n_{1,\ell,m}^{\left(z_{\ell,m,n}\right)}+\alpha_{\ell,z_{\ell,m,n}}-1\right)}{\sum_{k=1}^{K}\left(n_{1,\ell,m}^{(k)}+\alpha_{\ell,k}\right)-1}\dfrac{n_{\ell,m}^{(1)}+\gamma_1-1}{\sum_{s=0}^{1}\left(n_{\ell,m}^{(s)}+\gamma_s\right)-1}\dfrac{n_{\ell,z_{\ell,m,n}}^{\left(w_{\ell,m,n}\right)}+\beta_{1,w_{\ell,m,n}}-1}{\sum_{v=1}^{V}\left(n_{\ell,z_{\ell,m,n}}^{(v)}+\beta_{1,v}\right)-1}, & \text{if} \quad x_{\ell,m,n}=1\end{cases} \tag{2}$$

Here, $\vec{z}_{\neg(\ell,m,n)}$ and $\vec{x}_{\neg(\ell,m,n)}$ represent the topic and preference status assignments for all word tokens except $w_{\ell,m,n}$, respectively. $n_{0,\ell,m}^{(k)}$ and $n_{1,\ell,m}^{(k)}$ denote respective number of tokens in the document m of the corpus ℓ that are assigned to common topic k and specific topic k.$n_{\ell,m}^{(s)}$ is the number of tokens with the preference status (0 or 1) in the document m of the corpus ℓ. $n_k^{(v)}$ is the number of the tokens of word v that are assigned to the common topic k, and $n_{\ell,\mathrm{k}}^{(v)}$ is the number of tokens of word v in the corpus ℓ that are assigned to topic k. Using the expectation of Dirichlet/Beta distribution, the model parameters in Table 3 can be readily obtained as follows:

$$\theta_{\ell,m,k}=\frac{n_{0,\ell,m}^{(k)}+\alpha_{0,k}}{\sum_{k=1}^{K}\left(n_{0,\ell,m}^{(k)}+\alpha_{0,k}\right)} \tag{3}$$

$$\vartheta_{\ell,m,k}=\frac{n_{1,\ell,m}^{(k)}+\alpha_{\ell,k}}{\sum_{k=1}^{K_\ell}\left(n_{1,\ell,m}^{(k)}+\alpha_{\ell,k}\right)} \tag{4}$$

$$\phi_{k,v}=\frac{n_k^{(v)}+\beta_{0,v}}{\sum_{v=1}^{V}\left(n_k^{(v)}+\beta_{0,v}\right)} \tag{5}$$

$$\varphi_{\ell,k,v}=\frac{n_{\ell,k}^{(v)}+\beta_{1,v}}{\sum_{v=1}^{V}\left(n_{\ell,k}^{(v)}+\beta_{1,v}\right)} \tag{6}$$

$$\lambda_{\ell,m,s}=\frac{n_{\ell,m}^{(s)}+\gamma_s}{\sum_{v=1}^{V}\left(n_{\ell,m}^{(s)}+\gamma_s\right)} \tag{7}$$

In our case, $L = 2$ denotes two corpora for scientific publications and patents. The symmetric Dirichlet/Beta priors $\alpha_0, \alpha_\ell, \beta_0, \beta_1$ and γ are set at 0.1, 0.1, 0.001, 0.001 and 0.5, respectively. The Gibbs sampling is run for 1000 iterations, including 200 for the burn-in period.

Experimental results and discussions

Dataset

The pharmaceutical field is rich in information resources of science, technology and drugs, and the linkages between science and technology are prominent (Glänzel & Meyer, 2003; Narin et al., 1997). Our dataset is derived from a comprehensive, freely accessible online database in this field, DrugBank.[1] Due to intensive science-based innovation embodied in the drugs, related scientific publications and patents are explicitly linked to the resulting drugs. This provides an opportunity for this study to detect commonalities and specialties between scientific publications and patents from a same field. Though several conclusions are drawn from two independent chemical compounds corpora in Xu et al. (2019a), different search strategies for two corpora may result in biased conclusions. To say it in another way, the results in this study should be more credible than prior works.

The DrugBank dataset is downloaded on 1st November, 2019 in the XML format, and parsed to MySQL database. Then, the titles and abstracts of scholarly articles are fetched from PubMed database with E-Fetch API,[2] and those of patent documents are retrieved from EPO database with OPS API.[3] Finally, the number of scientific publications and patents is 10,355 and 5932 respectively, but only 9357 scholarly articles and 5817 patent documents are attached the corresponding abstract information. This study detects the sentences in the titles and abstracts with *geniass* (Sætre et al., 2007) and tokenizes the segmented sentences with *geniatagger* (Tsuruoka et al., 2005). Similar to Xu et al. (2019b), all numbers are replaced with a special word *NUMBER*. In order to discover the language characteristics of scientific publications and patent documents from the perspective of syntactic and lexical complexity, the Stanford CoreNLP toolkit[4] is utilized for Part-of-Speech (PoS) tagging and parsing, and Tregex[5] for extracting the number of clauses. Note that, before calculating the lexical complexity, several PoS tags with similar meaning are merged. For example, "NN", "NNS", "NNP" and "NNPS" are viewed as nouns here. In addition, to reduce the interference of un-related information, copyright information is removed with the rules in Xu et al. (2021).

In addition, to highlight linguistic characteristics of scientific publications and patents, the Wikipedia gold standard (WikiGold for short) corpus (Balasuriya et al., 2009) serves as a proxy of the generic texts. This corpus was originally created to evaluate the performance of named entity recognition models. There are 145 articles in this corpus in total, which were selected at random from all articles describing named entities in the May 22,

[1] https://www.drugbank.ca/.
[2] https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch.
[3] http://ops.epo.org/.
[4] https://stanfordnlp.github.io/CoreNLP/index.html.
[5] https://nlp.stanford.edu/software/tregex.shtml.

Table 4 Statistics of word tokens and unique words in scientific publications, patents, and WikiGold copora

	Original corpus	Filtered corpus	Stopwords	Rate (%)
Scientific publications				
Word tokens	2,121,177	501,785	1,619,392	76.34
Unique words	42,252	37,252	13,237	11.83
Patents				
Word tokens	565,538	93,949	471,589	83.39
Unique words	12,705	10,999	6392	13.43
WikiGold				
Word tokens	39,007	6812	32,195	82.54
Unique words	7443	2739	5219	63.20

2008 dump of English Wikipedia. For more details on this corpus, we refer the readers to Balasuriya et al. (2009).

Stopword identification

Descriptive statistical analysis

After running the HMM-LDA model, each word token is assigned a category which indicates syntactic function word or semantic function word. From the HMM-LDA results, Table 4 reports the number of word tokens and unique words before and after filtering stopwords in scientific publications, patents and WikiGold corpora. One can see that nearly 80% word tokens are removed from these three corpora, which is consistent with the observations in Gerlach et al. (2019). However, reduction rates for unique words are 11.83%, 13.43% and 63.20% for scientific publications, patents and WikiGold corpora, respectively. This is not in line with the observations in Gerlach et al. (2019). In our opinion, main reason is that the HMM-LDA model is able to handle ambiguous terms very well, but the information theoretic approach (Gerlach et al., 2019) cannot. It is worth mentioning that there seems no significant difference between S&T literature and generic texts in term of rate of word tokens, but a different pattern between them can be observed in term of rate of unique words.

As we all know, the purposes of scientific publications and patents are different from each other. The purpose of publishing scientific papers is to spread and share knowledge, and the focus of applying for patents is to protect intellectual property. This enables different statements to appear in the scientific publications and patents. Xu et al. (2019a) observed that the intersection of unique words is lower than 30%. However, as for the DrugBank dataset, the overlapping syntactic function and semantic function words account for one fifth and more than half of unique words for the scientific publications and patent documents respectively (25.86% vs. 18.92% and 53.55% vs. 64.08%). This indicates that a large part of words used in the patents are also shared by scientific publications from a same domain. Further, the nouns dominate the overlapping words, as shown in Table 5.

To have an intuitive understanding, Table 6 shows a comparison of semantic function and syntactic function words specific to scientific publications, patents and the overlapping between them. From Table 6, it is not difficult to see that the HMM-LDA model recognizes common stopwords such as "*a*", "*and*" and "*the*" in the text, and also identifies background

Table 5 Statistics of intersection of unique words between scientific publications and patents

	Original corpus	Filtered corpus	Stopwords
Noun	5528	4727	1687
Verb	1365	863	755
Adj	1311	1043	617
Adv	317	227	206
Others	231	188	158
Σ	8752	7048	3423

Table 6 A list of several semantic function and syntactic function words for scientific publications, patents and the overlapping between them

Semantic function words			Syntactic function words		
Scientific publications	Patents	Overlap	Scientific publications	Patents	Overlap
NUMBER	NUMBER	Drug	of	the	of
receptor	alkyl	Therapy	NUMBER	a	and
cancer	hydrogen	Cancer	treatment	NUMBER	are
plasma	treatment	Plasma	clinical	to	treatment
treatment	cancer	Receptor	effect	invention	method

function words such as "*effect*", "*method*" and "*treatment*" that do not contribute to thematic structure. There are a certain number of overlapping semantic function words between scientific publications and patents, such as "*receptor*", "*cancer*" and "*plasma*" and syntactic function words such as "*of*", "*and*" and "*treatment*". In the meanwhile, several ambiguous words can also be observed, such as "*NUMBER*" and "*treatment*", and corpus-specific syntactic function words, such as "*clinical*" and "*invention*".

Comparison with TF-IDF based heuristic method

In order to further demonstrate the advantages of our stopword identification method, we benchmark our approach against TF-IDF based heuristic method. In fact, it is not trivial to evaluate directly the performance of stopword identification approach, especially when the *ground truth* is unavailable. Gerlach et al. (2019) put forward an indirect evaluation approach by quantifying how much the inferred thematic structures correlated with category labels from the document metadata.

Thus, category labels are required in advance for scholarly articles and patent documents. In this work, *Web-of-Science-Categories* of each scholarly article are retrieved from Web of Science according to the DOI name (Xu et al., 2019c). As for the patents, 4-character IPC codes are used as category labels. Then, for simplicity, the LDA model is utilized in this subsection to extract thematic structures from filtered corpora. Finally, the normalized mutual information (NMI) (Xu et al., 2016, 2018) and adjusted mutual information (AMI) (Xu et al., 2016, 2018) are calculated between discovered themes and category labels, as illustrated in Fig. 5.

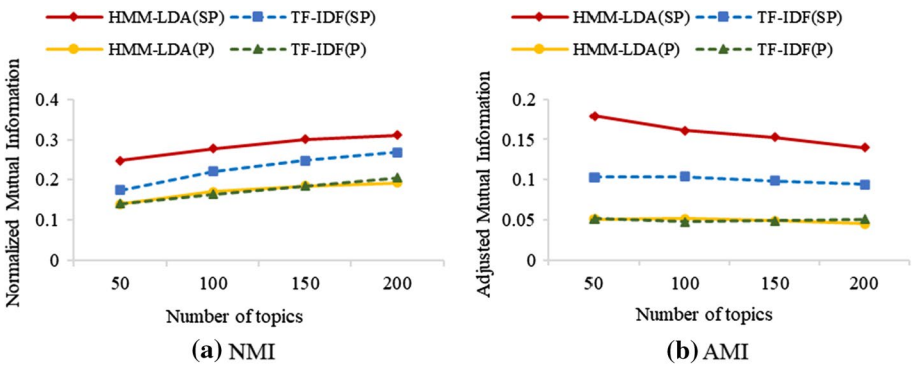

Fig. 5 Performance of the HMM-LDA model and TF-IDF based heuristic method on scientific publications (SP) and patents (P) in terms of NMI (**a**) and AMI (**b**)

Note that the following parameters are used in this comparative analysis. In the LDA model, the symmetric Dirichlet priors α and β are set at 0.5 and 0.01 respectively. The Gibbs sampling is run for 1000 iterations, including 200 for the burn-in period. The number of topics assumes from the set {50, 100, 150, 200}. As for the TF-IDF based heuristic method, similar to Gerlach et al. (2019), the threshold is fixed to 9. The main reasons for this setting are twofold: (1) the rate of unique words and tokens between our corpora and that in Gerlach et al. (2019) is similar; (2) the desired reduction of the data size is also similar (i.e., more than 80%).

From Fig. 5, one can see that the HMM-LDA model obviously outperforms the TD-IDF based heuristic method on the scientific publications, and similar performance can be observed on the patents in terms of the NMI and AMI. Furthermore, the performance difference between these two approaches seems be independent from the number of topics. As a whole, the HMM-LDA model is superior to the TF-IDF based heuristic method.

Analysis of linguistic characteristics

In this subsection, we analyze linguistic characteristics of scientific publications and patents. To check the influence of stopwords on linguistic characteristics, two groups of indicators on original corpora and filtered versions are reported here.

Syntactic complexity

Length and sentence length

For length and sentence length indicators, we average the number of word tokens in the resulting texts (cf. Table 2). From the first rows in Table 7, we can observe the different English writing style followed by general English corpus and S&T texts. More specifically, the latter tends to use longer sentences. In addition, it can be seen scientific publications contain more word tokens than those of patents, regardless of whether stopwords are filtered out.

As we all know, the abstract length (AL) of scientific publications and patents may be affected by the submission guideline and patent guide. To validate this point, we track the

Table 7 The linguistic characteristics of scientific publications and patents in terms of syntactic complexity indicators

	WikiGold corpus			Scientific publications		Patents	
	Original corpus	Filtered corpus		Original corpus	Filtered corpus	Original corpus	Filtered corpus
L	269.014	46.979	TL	14.652	5.165	9.270	1.542
			AL	210.480	47.911	87.768	14.578
SL	21.326	4.024	ASL	29.669	6.858	47.225	7.609
SC	1.183	–	TSC	0.111	–	0.022	–
			ASC	1.214	–	1.524	–

journal submission guidelines of several top journals with the number of articles greater than 50 in our dataset. From Table 11 in the "Appendix", most journals limit the abstract length not to be greater than 250 words. Similarly, we also check the USPTO patent guide and a patent document must contain a concise summary in the abstract (not be longer than 150 words). These specifications should be main reason that scientific publications have longer abstracts than patents.

From the perspective of sentence length of abstract (ASL), scientific publications are similar to previous observation in Lu et al. (2019). But the sentence length of patent documents is greater than those of scholarly articles and generic texts. It is very possible for the inventors to use longer sentences to explain their inventions in order to ensure the novelty, creativity and patentability.

In addition, by comparing the indicators' values of scientific publications and patents before and after filtering out the stopwords, we can see that the reductions of patents' indicators are greater than those for scientific publications. This indicates again that the patents contain more stopwords.

Sentence complexity

The last two rows in Table 7 (SC, TSC and ASC) illustrate the average number of clauses in each sentence for three corpora. S&T texts use more clauses than generic texts. Scientific publications include more clauses in the titles, but patents tend to use more clauses in the abstracts.

Lexical complexity

Lexical diversity

This indicator measures the total number of unique words normalized by the length of the resulting text (cf. Table 2). From Table 8, it can be seen that filtered corpus shows more significant characteristics than its original one in terms of TDIV and ADIV. This indicates that the stopwords are mentioned multiple times in the texts to conform to a certain language specification. This is in line with our intuition. Among these three corpora, the patent documents use more abundant words, followed by generic texts.

Table 8 The linguistic characteristics of scientific publications and patents in terms of lexical complexity indicators

	WikiGold corpus			Scientific publications		Patents	
	Original corpus	Filtered corpus		Original corpus	Filtered corpus	Original corpus	Filtered corpus
DIV	0.595	0.695	TDIV	0.939	0.978	0.951	0.991
			ADIV	0.568	0.670	0.629	0.857
DEN			TDEN				
Noun	0.373	0.837	Noun	0.488	0.827	0.553	0.811
			Verb	0.036	0.022	0.075	0.044
Verb	0.109	0.043	Adj	0.113	0.125	0.123	0.117
			Adv	0.008	0.006	0.013	0.009
			ADEN				
Adj	0.056	0.089	Noun	0.362	0.801	0.332	0.698
			Verb	0.103	0.033	0.115	0.095
Adv	0.027	0.005	Adj	0.089	0.120	0.093	0.122
			Adv	0.030	0.012	0.033	0.027
SOP			TSOP				
Noun	6.604	6.764	Noun	8.204	8.301	8.382	8.439
			Verb	7.877	8.427	8.714	8.894
Verb	5.060	7.903	Adj	8.305	8.731	9.033	9.102
			Adv	6.193	6.306	7.567	6.957
			ASOP				
Adj	7.034	7.307	Noun	7.087	7.321	7.382	7.436
			Verb	6.338	8.357	6.999	8.373
Adv	5.591	7.490	Adj	7.835	8.603	7.889	8.328
			Adv	6.651	7.876	7.731	7.987

Lexical density

In order to further discover the characteristics of lexical complexity at a finer granularity, we consider the average ratio of word tokens with the following parts of speech: noun, verb, adj. and adv. (cf. Table 2) As shown in Table 8, for original corpus of abstracts and generic texts, a common pattern of three corpora can be observed: the nouns appear most frequently, followed by verbs, and adverbs appear least. This observation is in accord with Lu et al. (2019). After filtering out the stopwords, a similar pattern is also observed, but the adjectives rank second. That is to say, a large proportion of verbs belong to syntactic function words.

Comparing the indicators of scientific publications and patents among original and filtered corpora, we find more interesting results. A significant difference is the change in the proportion of adjectives in the title part. This indicates that the adjectives appearing in the titles of scientific publications are more beneficial to express semantic structures than those in the patents. But the nouns appearing in the abstracts of scientific publications are denser.

Lexical sophistication

This indicator measures the average length of word tokens with the following parts of speech: noun, verb, adj. and adv. (cf. Table 2) From Table 8, one finding is that the terms from scientific publications and patents are longer than those in generic texts. The average length of adjectives is longest in original versions of three corpora. This observation is again in accord with Lu et al. (2019). In addition, one notable change is that the length of most part-of-speech words has become greater after filtering stopwords. In other words, most semantic function words consist of more characters.

Similarly, we compare the indicators between original and filtered versions. The word tokens used in the patents are usually longer than those in scholarly articles. Among these four categories of words, the adjectives rank first in term of average length in the original corpora, and adverbs and verbs rank first from last in the original titles and abstracts respectively. As for the filtered corpora, the verbs rank first in the patents and generic texts.

General remarks

In summary, several linguistic characteristics are shared between scientific publications and patents as follows: (1) Compared to semantic function words, the stopwords are mentioned multiple times in the texts; (2) The nouns are used most frequently; (3) Most verbs are syntactic function words which do not contribute to thematic structures; (4) The adjectives rank first in term of average length in original corpora; (5) The semantic function words mainly consist of longer terms.

Though, the two-tailed independent-samples *t* test with 95% confidence interval indicate that there exists statistically significant difference between scientific publications and patents in terms of all syntactic complexity indicators, and in terms of nearly 80% lexical complexity indicators (cf. Table 12 in the "Appendix"). In what follows, we can also summarize the customized linguistic characteristics between scientific publications and patents. (1) The scientific publications contain more word tokens than patents, and the titles in scholarly articles often use more clauses, while patent documents have usually longer sentence and use more clauses in the abstracts. (2) The patents contain more synthetic function words than scientific publications. (3) The nouns appear more frequently in the abstracts of academic articles than patents. (4) The word tokens in the patents are usually longer than those in scholarly articles. (5) The verbs rank first in the filtered abstracts of patents, and the adjectives first in the filtered abstracts of scientific publications.

Multi-collection topic analysis

Number of topics

For the sake of identifying a proper number of common and special topics, the perplexity is calculated for each candidate value combination of the number of common topics K, the number of topics specific to scientific publications K_1, and the number of

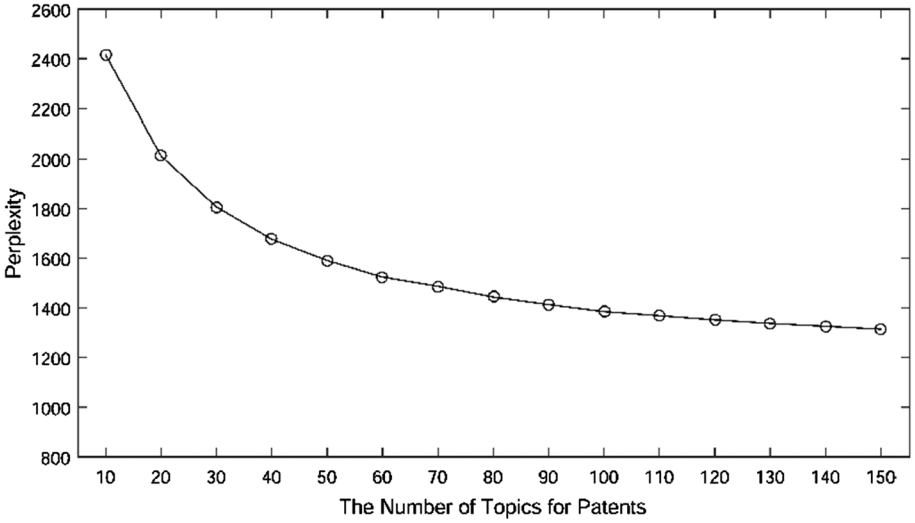

Fig. 6 The perplexity with different number of topics

topics specific to patents K_2. As a standard measure for model selection, this measure is defined as the exponential of the negative normalized predictive likelihood under the model $\mathcal{M}$ [cf. Equation (8)], and a lower value indicates a better modeling performance.

$$\Pr\left(\vec{\tilde{w}}|\mathcal{M}\right) = \exp - \frac{\sum_{\ell=1}^{L}\sum_{m=1}^{M_\ell} \log P\left(\vec{\tilde{w}}_{\ell,m,\cdot}|\mathcal{M}\right)}{\sum_{\ell=1}^{L}\sum_{m=1}^{M_\ell} N_{\ell,m}} \tag{8}$$

Here, the likelihood of a text document of the test corpus $\Pr\left(\vec{\tilde{w}}_{\ell,m,\cdot}|\mathcal{M}\right)$ can be directly expressed as a function of the multinomial parameters as follows.

$$\Pr\left(\vec{\tilde{w}}_{\ell,m,\cdot}|\mathcal{M}\right) = \prod_{n:x_{\ell,m,n}=0}\sum_{k=1}^{K} \phi_{k,\tilde{w}_{\ell,m,n}}\tilde{\theta}_{\ell,m,k}\tilde{\lambda}_{\ell,m,0} \times \prod_{n:x_{\ell,m,n}=1}\sum_{k=1}^{K_\ell} \varphi_{\ell,k,\tilde{w}_{\ell,m,n}}\tilde{\vartheta}_{\ell,m,k}\tilde{\lambda}_{\ell,m,1} \tag{9}$$

According to Kim et al. (2015), a better performance can be obtained when the values of topic number are set closer to the real-world cases. In the DrugBank dataset, each drug is attached with multiple ATC (Anatomical Therapeutic Chemical) codes. The ATC classification is an internationally accepted classification system for medicines which is maintained by the World Health Organization (WHO). This classification system consists of five levels, and the first level has 14 unique codes. Hence, the number of common topics K is fixed to 14 in this study. It has been shown that there exists the positive correlation between the number of documents and the number of topics (Schmiedel et al., 2019). Since the number of academic articles is about 1.5 times of the number of patent documents (cf. "Dataset" section), the number of topics for scientific publications is set to 1.5 times of the number of topics for patents, viz. $K_1 = 1.5K_2$. This work lets the candidates for the number of topics for patents K_2 from 10 to 150 with a step size 10. Figure 6 depicts the perplexity with different number of topics. From Fig. 6, it is not difficult to see that the perplexity of the CDTM model converges when the number of topics for patents K_2 is 100, so the

Table 9 An illustration of 2 common topics and 4 special topics for the DrugBank dataset

Topic C1		Topic C8	
Common topics			
Patients	0.031	disease	0.011
Blood	0.016	gastric	0.010
Heart	0.014	contrast	0.008
Hypertension	0.010	imaging	0.008
Cardiac	0.008	acid	0.007
Cox	0.007	therapy	0.007
Coronary	0.007	patient	0.007
Platelet	0.007	agent	0.006
Angiotensin	0.006	gastrointestinal	0.006
Myocardial	0.006	conditions	0.006

Scientific publication				Patent			
Topic S71		Topic S102		Topic T80		Topic T90	
Special topics							
gastric	0.072	egfr	0.073	particles	0.010	derivatives	0.140
acid	0.053	mutant	0.044	applications	0.041	imaging	0.102
reflux	0.037	growth	0.039	mucosal	0.041	radiolabeled	0.021
ulcer	0.034	kinase	0.026	polymeric	0.037	mole	0.019
omeprazole	0.029	receptor	0.031	transport	0.037	styrylpyridine	0.016
lansoprazole	0.028	tyrosine	0.026	surface	0.033	quality	0.016
ranitidine	0.020	gene	0.020	carriers	0.026	transfer	0.013
gastroesophageal	0.016	lung	0.017	density	0.024	hydralazine	0.013
peptic	0.016	tki	0.013	solubility	0.024	alkylated	0.013
duodenal	0.015	cell	0.012	mucus	0.015	htla	0.013

Each topic is shown with top 10 words and their corresponding probability conditioned on that topic

number of topics for scientific publications and patents are fixed respectively to 150 and 100 in this work.

Common and special topics

In this subsection, we discuss the interesting discoveries by applying our revised CDTM model to the DrugBank dataset consisting of scientific publications and patents. Table 9 lists 2 common topics and 4 special topics from this dataset, in which each topic is shown with top 10 words and their corresponding probability conditioned on that topic.

The common topics reflect thematic structures shared by scientific publications and patents. One can see that the terms for common topics are mainly more general descriptive words in the pharmaceutical field, such as “patients” and “blood” in Topic C1 and “disease” and “therapy” in Topic C8. As for the special topics, the top-ranked words for scientific publications and patents clearly reveal their characteristics. Scientific publications tend to represent the description of the disease mechanism and the medication content, such as “gastroesophageal” and “omeprazole” in Topic S71, and “growth” and “gene” in Topic S102. However, the themes from patents are biased towards the preparation and practical

Table 10 Thematic structures of several scientific publications and patents

Type	Title	Theme distribution
I: Special	Atazanavir: a review of its use in the management of HIV-1 infection [PMID:19496633]	S5: 71.16%, S98: 4.49%, S44: 1.59%, S64: 1.59%
	Abstract LB-100: discovery of HM61713 as an orally available and mutant EGFR selective inhibitor [https://doi.org/10.1158/1538-7445.AM2014-LB-100]	S102: 68.11%, S14: 8.28%, S32: 2.54%
	RNA interference mediating small RNA molecules [PN:US8372968]	T17: 82.62%, T66: 2.63%, T6: 1.38%, T93: 1.38%
	Method for delivering a pharmaceutical composition to patient in need thereof [PN:US9393208]	T80: 65.35%, T90: 2.56%, T11: 2.56%
II: Common	Oral factor Xa inhibitors for thromboprophylaxis in major orthopedic surgery: a review [PMID:19696978]	C1: 94.93%, C3: 2.42%
	Shortcomings of the first-generation proton pump inhibitors. [PMID:11430506]	C8: 81.31%, C2: 2.48%, C7: 2.48%
	Platelet aggregation inhibition using low molecular weight heparin in combination with a GP IIb/IIIa antagonist [PN:US6136794]	C1: 86.78%, C8: 6.32%
	Methods and compositions for the treatment of gastrointestinal disorders [PN:US8110553]	C8: 93.70%, C1: 2.10%, C12: 2.10%
III: Hybrid	Third-generation inhibitors targeting EGFR T790M mutation in advanced non-small cell lung cancer [PMID: 27071706]	C12: 70.45%, S102: 64.23%
	Clinical pharmacokinetics and pharmacodynamics of insulin glulisine. [PMID: 18076215]	C9: 75.53%, S109: 70.80%, C2: 11.70%
	De-agglomerator for breath-actuated dry powder inhaler [PN:US6871646]	C2: 87.50%, T84: 71.50%
	Compositions containing alpha-2-adrenergic agonist components [PN:US9687443]	C6: 73.78%, T21: 42.63%, C2: 18.90%

application of drugs, such as "particles" and "solubility" in Topic T80, and "mole" and "styrylpyridine" in Topic T90.

In order to further verify whether the discovered themes make sense, we sample a couple of documents and check their titles and corresponding topics, as shown in Table 10. According to theme distributions, the documents can be divided into the following three categories:

- *Type I: Special* The documents in this type have high probability distribution on their own special topics. For example, scientific publication (PMID: 19496633) mainly discussed the theme (S102) of treatment approaches for lung cancer. Activating mutations of EGFR are well known as oncogenic driver mutations in lung adenocarcinoma (Lee et al., 2014). Patent (PN: US9393208) mainly focus on the topic (T80) of method for delivering pharmaceutical compositions, which demonstrates the practical application of drugs.
- *Type II: Common* This type of documents distributes on the common topics with a large probability. Let's take common topic C8 (gastrointestinal diseases therapy) as an example. Our revised CDTM model assigns this theme with the probability 93.70% and 81.31% to the patent (PN: US8110553), and scholarly article (PMID: 11430506), respectively. Proton pump inhibitors (PPIs) are widely prescribed for the treatment of gastro-oesophageal reflux disease (GORD) as well as gastric and duodenal ulcers (Tytgat, 2001). From the resulting titles, it is not difficult to see that our theme assignments are rational.
- *Type III: Hybrid* This group of documents involves both common topics and their special topics with a high probability. For instance, scientific publication (PMID: 27071706) is about common topic C12 (medication of lung cancer) with the probability 70.45% and scientific topic S102 (treatment approaches for lung cancer) with the probability 64.23%. The patent (PN: US6871646) mainly investigated common topic C2 (dry powder inhaler) with the proportion 87.50% and technological topic T84 (medical equipment) with the probability 71.50%.

To obtain insights on the relations amongst common and special topics, Fig. 7 visually illustrates the cosine similarity between thematic structures (viz., $\vec{\phi}_k$ and $\vec{\varphi}_{\ell,k}$ in Fig. 4). Each node is sized by the topic proportion (viz., $\vec{\theta}_{\ell,m}$ and $\vec{\vartheta}_{\ell,m}$ in Fig. 4). Different color in Fig. 7 indicates different cluster. These clusters are obtained with *VOSviewer* (van Eck & Waltman, 2010), where the Fractionalization method is used for normalizing the strength of the edges between nodes, and the attraction and repulsion are set to 4 and -2. As shown in Fig. 7, common topics of scientific publications and patents in the DrugBank dataset are mainly concentrated in the middle of the network, and special topics at the periphery of the network. The important common and special themes structures in the collection respectively reflect the commonalities and specialties of scientific publications and patents.

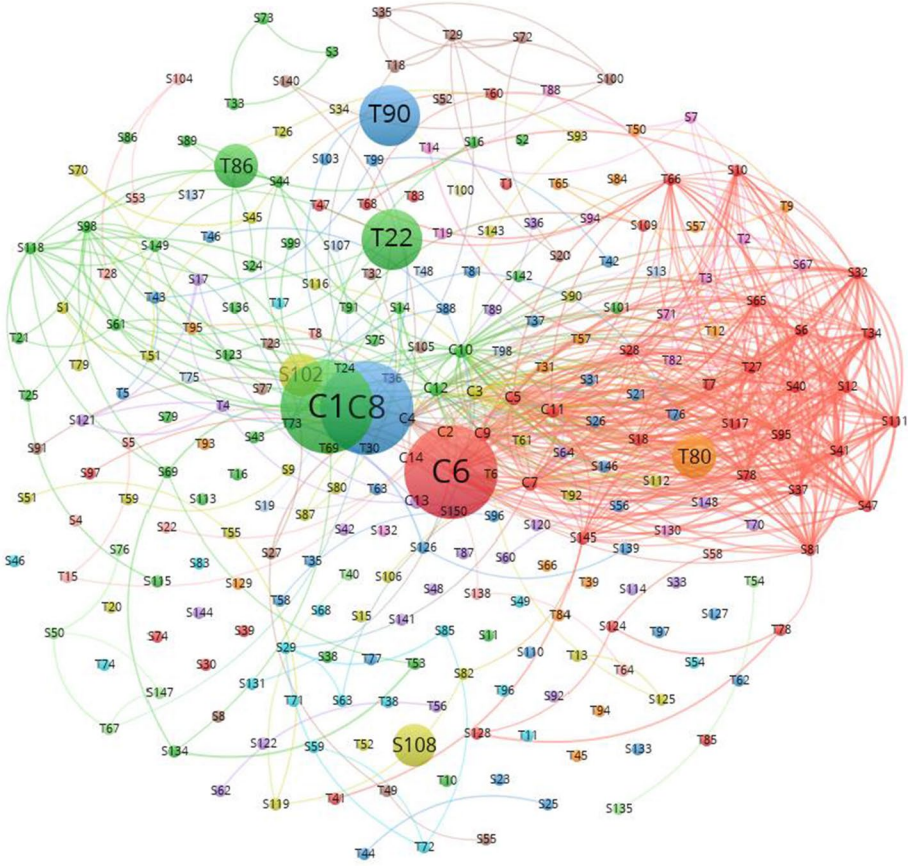

Fig. 7 The connections amongst common topics and special topics in the DrugBank dataset

Conclusions

The intersection and integration of science and technology in modern society has become an important driving force for technologies to emerge. Due to different purpose, statement and quality of scientific publications and patents, science-technology topic linkages meet some challenges. In order to understand the difficulties and improve the performance, a framework is developed here to detect the commonalities and specialties between scientific publications and patents from two perspectives: linguistic characteristics and thematic structures. In more details, our detection framework integrates syntactic and lexical complexity indicators and a revised joint thematic structure discovery

model for multiple corpora as well as an advanced stopword identification method on the basis of HMM-LDA model.

Extensive experiments are conducted on the DrugBank dataset. Several conclusions can be drawn as follows: Five commonness and five significant differences in terms of linguistic characteristics. For example, nouns are used most frequently among them, and scientific publications contain more word tokens than patent documents, but patents have usually longer sentence and use more clauses. As for thematic structures, the topics about general description in the pharmaceutical field are shared by scientific publications and patents. Different emphasis or characteristics are also observed for these two different corpora. The scientific publications tend to explain the disease mechanism and the medication content, while the patents bias towards the preparation and practical application of drugs.

The results of our study have brought several detailed information at the lexical level for the linkage research between science and technology. Though, several limitations are also identified in this study. Academic articles and patent documents contain many entity mentions, such as drug name, family name and so on. It is far from mature to recognize accurately these domain entity mentions, since the patterns of linguistic characteristics in academic publications and patent documents are very different from generic texts (Chen et al., 2020; Xu et al., 2015). In the near future, the cutting-edge NLP techniques, such as deep learning model, will be adopted for entity mention identification. In the meanwhile, other indicators such as readability (Gazni, 2011; Hartley et al., 2003) will be further considered in our next work. In addition, the validity of our framework and the consistency of results need to be established through further testing on diverse fields of science and technology.

Appendix

See Tables 11 and 12.

Table 11 Journals and abstract restrictions corresponding to the top 20% papers that appear most frequently

Journal	Abstract words limitation restrictions
Drugs	150 to 250 words
Antimicrobial Agents and Chemotherapy	250 words or fewer
New England Journal of Medicine	250 words or fewer
Journal of Pharmacology and Experimental Therapeutics	250 words or fewer
Clinical Pharmacokinetics	150 to 250 words
Drug Metabolism and Disposition	250 words or fewer
Expert Opinion on Pharmacotherapy	200 words or fewer
Veterinary Record	200 words or fewer for research papers or scientific reviews, and 150 words or fewer for short communications
British Journal of Clinical Pharmacology	250 words or fewer for original articles, review articles, systematic reviews and (network) meta-analysis, and 150 words or fewer for short reports
Lancet	300 words or fewer
Clinical Cancer Research	250 words or fewer for clinical trial brief reports, research articles and clinical trial brief reports, and 150 words or fewer for CCR reviews, perspectives, perspectives on regulatory science and policy
British Journal of Pharmacology	250 words or fewer for research papers, and 150 words or fewer for review articles, invited mini-reviews and hypothesis articles
American Journal of Veterinary Research	250 words or fewer
European Journal of Pharmacology	250 words or fewer
The Cochrane database of Systematic Reviews	Not clearly stated
Journal of Biological Chemistry	250 words or fewer
Journal of Clinical Pharmacology	250 words or fewer
Clinical Therapeutics	400 words or fewer
Proceedings of the National Academy of Sciences of the United States of America	250 words or fewer
Expert Opinion on Investigational Drugs	200 words or fewer
Journal of Veterinary Pharmacology and Therapeutics	200 words or fewer
Arzneimittelforschung	250 words or fewer
Blood	250 words or fewer
Annals of Pharmacotherapy	250 words or fewer

Table 12 Two-tailed independent sample t-test results of 11 language complexity indicators

	t	*df*	Sig. (two-tailed)	Lower	Upper
Syntactic complexity					
TL					
Original corpus	38.310	11,582.470	**.000*****	5.106	5.657
Filtered corpus	89.438	16,095.530	**.000*****	3.543	3.702
AL					
Original corpus	76.907	14,506.691	**.000*****	119.584	125.839
Filtered corpus	79.179	15,119.751	**.000*****	32.508	34.158
ASL					
Original corpus	− 26.284	6098.024	**.000*****	− 18.865	− 16.246
Filtered corpus	− 3.795	6285.743	**.000*****	− 1.139	− 0.363
TSC	23.935	15,502.260	**.000*****	0.082	0.097
ASC	− 14.237	6356.413	**.000*****	− 0.353	− 0.267
Lexical complexity					
TLD*i*					
Original corpus	− 6.752	9111.634	**.000*****	− 0.015	− 0.008
Filtered corpus	− 11.144	8490.105	**.000*****	− 0.015	− 0.011
ALD*i*					
Original corpus	− 23.245	11,876.567	**.000*****	− 0.066	− 0.056
Filtered corpus	− 65.167	14,880	**.000*****	− 0.1934	− 0.1821
TLD*e*					
Original corpus					
Noun	− 25.462	9413.928	**.000*****	− 0.070	− 0.060
Verb	− 27.092	8071.445	**.000*****	− 0.041	− 0.036
Adj	− 4.951	9008.773	**.000*****	− 0.014	− 0.006
Adv	− 8.051	8774.687	**.000*****	− 0.006	− 0.004
Filtered Corpus					
Noun	3.143	4428.559	.002*	0.006	0.025
Verb	− 9.008	3878.448	**.000*****	− 0.027	− 0.017
Adj	1.940	4534.474	.052	− 0.000	0.016
Adv	− 3.123	3827.016	.002*	− 0.006	− 0.001
ALD*e*					
Original corpus					
Noun	34.798	10,922.737	**.000*****	0.028	0.032
Verb	− 17.365	8761.490	**.000*****	− 0.014	− 0.011
Adj	− 5.377	10,116.578	**.000*****	− 0.005	− 0.002
Adv	− 8.017	9352.130	**.000*****	− 0.004	− 0.002
Filtered corpus					
Noun	34.438	7230.304	**.000*****	0.097	0.109
Verb	− 34.732	6197.195	**.000*****	− 0.066	− 0.059
Adj	− 1.050	7950.687	.294	− 0.007	0.002
Adv	− 14.883	6255.715	**.000*****	− 0.017	− 0.013
TLS					
Original corpus					
Noun	− 6.116	11,079.242	**.000*****	− 0.235	− 0.121

Table 12 (continued)

	t	*df*	Sig. (two-tailed)	Lower	Upper
Verb	− 16.861	6289.824	**.000*****	− 0.934	− 0.739
Adj	− 12.361	5830.596	**.000*****	− 0.844	− 0.613
Adv	− 9.848	1750	**.000*****	− 1.648	− 1.100
Filtered corpus					
Noun	− 2.567	4317.907	.010	− 0.244	− 0.033
Verb	− 3.645	1682	**.000*****	− 0.718	− 0.216
Adj	− 3.658	5826	**.000*****	− 0.569	− 0.172
Adv	− 1.837	446	.067	− 1.348	0.045
ALS					
Original corpus					
Noun	− 17.061	10,042.968	**.000*****	− 0.329	− 0.261
Verb	− 33.907	10,138.161	**.000*****	− 0.699	− 0.623
Adj	− 2.148	9256.318	.032	− 0.104	− 0.005
Adv	− 24.550	7093.515	**.000*****	− 1.167	− 0.994
Filtered corpus					
Noun	− 3.874	8083.550	**.000*****	− 0.173	− 0.057
Verb	− 0.392	7077.579	.695	− 0.098	0.065
Adj	5.593	5223.721	**.000*****	0.178	0.370
Adv	− 1.100	4904	.271	− 0.310	0.087

The two rightmost columns are the lower and upper bounds of 95% C.I of the difference. Significant results are bolded

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Acknowledgements This work was supported partially by the National Natural Science Foundation of China (Grant Numbers 72074014 and 72004012). Our gratitude also goes to the anonymous reviewers and the editor for their valuable comments.

References

Albert, T. (2016). *Measuring technology maturity: Operationalizing information from patents, scientific publications and the web*. Springer.

An, X., Li, J., Xu, S., Chen, L., & Sun, W. (2021). An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics, 15*(2), 101135.

An, X., Xu, S., Wen, Y., & Hu, M. (2014). A shared interest discovery model for coauthor relationship in SNS. *International Journal of Distributed Sensor Networks, 2014*, 1–9.

Andy, S. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 15*, 707–719.

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., & Curran, J. (2009). Named entity recognition in Wikipedia. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources* (People's Web) (pp. 10–18). Suntec, Singapore.

Bassecouolard, E., & Zitt, M. (2004). Patents and publications: The lexical connection. In H. F. Moed, W. Glänzel, & U. Schoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 665–694). Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84.

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Brants, T. (2000). TnT: A statistical part-of-speech tagger. In *Proceedings of the sixth conference on applied natural language processing* (pp. 224–231). Somerset: ACL.

Brooks, H. (1994). The relationship between science and technology. *Research Policy, 23*(5), 477–486.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*, 263–311.

Calero-Medina, C., & Noyons, E. C. M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics, 2*(4), 272–279.

Chen, C., Buntine, W., Ding, N., Xie, L., & Du, L. (2015). Differential topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*, 230–242.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. *Scientometrics, 125*(1), 289–312.

Christopher, F. (1989). A stop list for general text. *ACM SIGIR Forum, 24*, 19–21.

Dubaric, E., Giannoccaro, D., Bengtsson, R., & Ackermann, T. (2011). Patent data as indicators of wind power technology development. *World Patent Information, 33*(2), 144–149.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*, 414–420.

Forti, E., Sobrero, M., & Franzoni, C. (2007). *The effect of patenting on the networks and connections of academic scientists* (pp. 272–284). Social Science Electronic Publishing.

Gao, H., Tang, S., Zhang, Y., Jiang, D., Wu, F., & Zhuang, Y. (2012b). Supervised cross-collection topic modeling. In *Proceedings of the 20th ACM international conference on multimedia* (pp. 957–960). New York: ACM.

Gao, J. P., Ding, K., Teng, L., & Pang, J. (2012a). Hybrid documents co-citation analysis: Making sense of the interaction between science and technology in technology diffusion. *Scientometrics, 93*, 459–471.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science, 37*, 273–281.

Gerard, S. (1963). Associative document retrieval techniques using bibliographic information. *ACM, 10*, 440–457.

Gerlach, M., Shi, H., & Amaral, L. A. N. (2019). A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence, 1*, 606–612.

Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations. *Scientometrics, 58*, 415–428.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. I*n Advances in neural information processing systems 17* (pp. 537–544). Vancouver, Canada.

Hartley, J., Pennebaker, J. W., & Fox, C. L. (2003). Abstracts, introductions and discussions: How far do they differ in style? *Scientometrics, 57*, 389–398.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the international ACM conference on research and development in information retrieval* (SIGIR'99) (pp.50–57). New York: ACM.

Hua, T., Lu, C.-T., Choo, J., & Reddy, C. K. (2020). Probabilistic topic modeling for comparative analysis of document collections. *ACM Transactions on Knowledge Discovery from Data, 14*, 24:1-24:27.

Huang, M. H., Yang, H. W., & Chen, D. Z. (2015). Increasing science and technology linkage in fuel cells: A cross citation analysis of papers and patents. *Journal of Informetrics, 9*, 237–249.

Kim, H., Choo, J., Kim, J., Reddy, C. K., & Park, H. (2015). Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the ACM international conference on knowledge discovery and data mining* (pp. 567–576). New York: ACM.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*, 148–161.

Lee, K., Mi, Y., Kim, M., Ji, Y., & Son, J. (2014). Abstract LB-100: Discovery of HM61713 as an orally available and mutant EGFR selective inhibitor. *Cancer Research, 74*(19 Supplement), LB-100.

Lee, M., Lee, S., Kim, J., Seo, D., Kim, P., Jung, H., Lee, J., Kim, T., Koo, H. K., & Sung, W. K., et al. (2011). Decision-making support service based on technology opportunity discovery model. In T.-H. Kim (Ed.), *FGIT-UNESST 2011* (Vol. 264, pp. 263–268). Springer.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., et al. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology, 70*, 462–475.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496.

Makrehchi, M., & Kamel, M. S. (2008). Automatic extraction of domain-specific stopwords from labeled documents. In *Proceedings of the 30th European conference on IR research* (pp. 222–233). Berlin: Springer.

Makrehchi, M., & Kamel, M. S. (2017). Extracting domain-specific stop words for text classifiers. *Intelligent Data Analysis, 21*, 39–62.

Montemurro, M. A., & Zanette, D. H. (2010). Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems, 13*, 135–153.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy, 26*, 317–330.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*, 492–518.

Paul, M. (2009). Cross-collection topic models: Automatically comparing and contrasting text. *Urbana, 51*, 61801.

Paul, M., & Girju, R. (2010). A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the 20th national conference on artificial intelligence* (pp. 545–550). CA: AAAI.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*, 257–286.

Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., & Ohta, T. (2007). AKANE system: protein-protein interaction pairs in the BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the 2nd BioCreative challenge evaluation workshop* (pp. 209–212). Madrid, Spain.

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation, 29*, 351–372.

Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods, 22*(4), 941–968.

Seki, K., & Mostafa, J. (2005). An application of text categorization methods to gene ontology annotation. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 138–145). New York: ACM.

Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Extracting the commercialization gap between science and technology—Case study of a solar cell. *Technological Forecasting and Social Change, 77*, 1147–1155.

Shibata, N., Kajikawa, Y., & Sakata, I. (2011). Detecting potential technological fronts by comparing scientific papers and patents. *Foresight, 13*, 51–60.

Takano, Y., Mejia, C., & Kajikawa, Y. (2016). Unconnected component inclusion technique for patent network analysis: Case study of internet of things-related technologies. *Journal of Informetrics, 10*(4), 967–980.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of the 10th Panhellenic conference on informatics* (pp. 382–382). Berlin: Springer.

Tytgat, G. (2001). Shortcomings of the first-generation proton pump inhibitors. *European Journal of Gastroenterology & Hepatology, 13*(Suppl 1), S29-33.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*, 523–538.

Verbeek, A., Debackere, K., & Luwel, M. (2002). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics, 54*, 399–420.

Wang, C., Thiesson, B., Meek, C., & Blei, D. (2009). Markov topic models. In *Proceedings of the 12th international conference on artificial intelligence and statistics* (pp. 583–590).

Wang, G., & Guan, J. (2011). Measuring science–technology interactions using patent citations and author-inventor links: An exploration analysis from Chinese nanotechnology. *Journal of Nanoparticle Research, 13*, 6245–6262.

Wang, Z., Xu, S., & Zhu, L. (2018). Semantic relation extraction aware of N-gram features from unstructured biomedical text. *Journal of Biomedical Informatics, 86*, 59–70.

Xu, H., Winnink, J., Yue, Z., Liu, Z., & Yuan, G. (2020). Topic-linked innovation paths in science and technology. *Journal of Informetrics, 14*(2), 101014.

Xu, S., An, X., Zhu, L., Zhang, Y., & Zhang, H. (2015). A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. *Journal of Cheminformatics, 7*(Suppl 1), S11.

Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019b). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*(4), 100983.

Xu, S., Hao, L., An, X., Zhai, D., & Pang, H. (2019c). Types of DOI errors of cited references in Web of Science with a cleaning method. *Scientometrics, 120*(3), 1427–1437.

Xu, S., Hao, L., Yang, G., Lu, K., & An, X. (2021). A topic models based framework for detecting and forecasting emerging technologies. *Technology Forecasting and Social Change, 162*, 120366.
Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., & Pang, H. (2018). Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. *Scientometrics, 117*(1), 61–84.
Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., & Li, L. (2016). Reviews on determining the number of clusters. *Applied Mathematics & Information Sciences, 10*(4), 1493–1520.
Xu, S., Zhai, D., Wang, F., An, X., Pang, H., & Sun, Y. (2019a). A novel method for topic linkages between scientific publications and patents. *Journal of the Association for Information Science and Technology, 70*(9), 1026–1042.
Xu, S., Zhu, L., Qiao, X., Shi, Q., & Gui, J. (2012). Topic linkages between papers and patents. In *Proceedings of the 4th international conference on advanced science and technology* (pp. 176–183).
Zhai, C., Velivelli, A., & Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 743–748). New York: ACM.
Zhang, H., Xu, S., & Qiao, X. (2014). Review on topic models integrating intra- and extra-features of scientific and technical literature. *Journal of the China Society for Scientific and Technical Information, 33*, 1108–1120.

Authors and Affiliations

Shuo Xu[1] · Ling Li[1] · Xin An[2] · Liyuan Hao[1] · Guancan Yang[3]

Shuo Xu
xushuo@bjut.edu.cn

Ling Li
infinitell@emails.bjut.edu.cn

Liyuan Hao
haoliyuan@emails.bjut.edu.cn

Guancan Yang
yanggc@ruc.edu.cn

[1] College of Economics and Management, Beijing University of Technology, Beijing 100124, People's Republic of China

[2] School of Economics and Management, Beijing Forestry University, Beijing 100083, People's Republic of China

[3] School of Information Resource Management, Renmin University of China, Beijing 100872, People's Republic of China