Connecting the Dots: Hypotheses Generation by Leveraging Semantic Shifts

Menasha Thilakaratne[(✉)], Katrina Falkner, and Thushari Atapattu

The University of Adelaide, Adelaide, South Australia, Australia
{menasha.thilakaratne,katrina.falkner,thushari.atapattu}@adelaide.edu.au

Abstract. Literature-based Discovery (LBD) (a.k.a. *Hypotheses Generation*) is a systematic knowledge discovery process that elicit novel inferences about previously unknown scientific knowledge by rationally connecting complementary and non-interactive literature. Prompt identification of such novel knowledge is beneficial not only for researchers but also for various other stakeholders such as universities, funding bodies and academic publishers. Almost all the prior LBD research suffer from two major limitations. Firstly, the over-reliance of domain-dependent resources which restrict the models' applicability to certain domains/problems. In this regard, we propose a *generalisable* LBD model that supports both *cross-domain* and *cross-lingual* knowledge discovery. The second persistent research deficiency is the mere focus of static snapshot of the corpus (i.e. ignoring the temporal evolution of topics) to detect the new knowledge. However, the knowledge in scientific literature changes dynamically and thus relying merely on static snapshot limits the model's ability in capturing semantically meaningful connections. As a result, we propose a novel *temporal* model that captures *semantic change* of topics using *diachronic word embeddings* to unravel more accurate connections. The model was evaluated using the largest available literature repository to demonstrate the efficiency of the proposed cues towards recommending novel knowledge.

Keywords: Literature-Based Discovery · Hypotheses generation · Diachronic word embeddings · Knowledge discovery · Literature mining

1 Introduction

Due to the massive influx of research publications, examining the published literature and constructing a novel research hypothesis in a sensible time-frame has almost become an unachievable endeavour for researchers. For example, consider a researcher who is interested in researching about *dementia*. To formulate a novel research hypothesis in the field, the researcher requires to comprehensively analyse and understand the existing body of knowledge. Currently, a simple search in *PubMed* alone for dementia results in more than 150,000 records.

Even though techniques such as text summarisation would assist the users to glean high-level overview of the field, they fail to elicit implicit interesting connections in disparate and seemingly independent facts that have the potential in developing novel knowledge.

To this end, *Literature-Based Discovery (LBD)* (a.k.a. *Hypotheses Generation*) which is a sub-discipline of text mining aims to infer such interesting cross-silo associations that bridge uncorrelated fragments of information to provide novel, actionable and meaningful insights in the field. For instance, consider two disjoint topics of interest (A and C); a therapeutic substance (e.g., *fish oil*) and a disease (e.g., *raynaud*) where the LBD process attempts to elicit novel *conceptual bridges* [3] (e.g., *blood viscosity*) that meaningfully connect the two knowledge fragments (Fig. 1). Hence, the ultimate motive of LBD research is to give new impetus to deduce new knowledge that will conclusively accelerate scientific productivity and research innovation. Discovering such conceptual bridges in a cross-disciplinary manner is the crux of the problem that we intend to address.
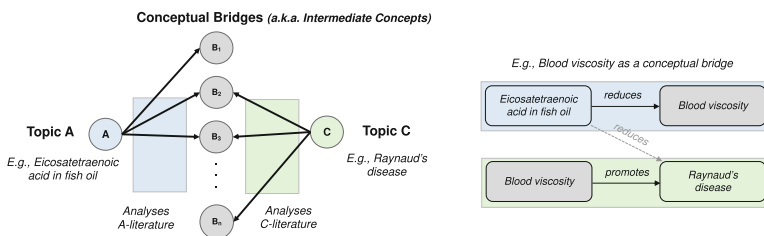

Fig. 1. Schematic overview of hypotheses generation

Despite the significant advances gained so far in the discipline, almost all the LBD systems suffer from two major drawbacks; 1) *Domain dependency:* the existing LBD studies are mainly restricted to the medical domain and rely on medical-related knowledge resources (e.g., *MeSH*, *UMLS*) *throughout* their workflow. Some of the LBD studies are not generalisable within the medical domain itself due to the usage of highly specialised resources [8]. As a result, extending these techniques in other non-medical LBD settings (e.g., *computer science* domain) is infeasible. Consequently, LBD research outside of the medical domain is in a nascent stage [8], and 2) *Static domains:* the prior LBD studies are based on the assumption that the domains remain static. This clearly hinders the model's performance in recommending time-aware novel knowledge linkages as the domains are changing dynamically, and new knowledge is being added to each domain every single day. Otherwise stated, the contribution of temporal cues in eliciting novel knowledge has been overlooked in the discipline [3].

To overcome the first limitation, a recent study of Sebastian et al. [7] has attempted to use *WordNet* in their LBD process. While this is encouraging, WordNet typically covers everyday English and limited in terms of scientific terminology. Considering the issues of domain-dependency in LBD studies arose the questions; *how to identify potential knowledge discovery cues whose success does*

not depend on domain-dependent resources such as MeSH and UMLS? and *what are the domain-independent resources that have a wider coverage than WordNet to perform the initial preprocessing of the literature?*

More recently, a few studies [3,4,12] have attempted to mitigate the second limitation, which is considering the domains to be static through the infusion of temporal dimension into the LBD process. Even though these studies undoubtedly invigorate the traditional LBD setting, they still contain several inherent limitations. Firstly, the temporal analysis component of these studies is fairly shallow. For example, Xun et al. [12] have only considered the first and last values of the temporal sequence to measure the trend of implicit associations by ignoring the patterns in the overall sequence. Secondly, as of most of the existing LBD literature, these studies rely on one or two temporal characteristics to discover potential new knowledge linkages. This is limiting as such methodologies may excessively be picking only one type of novel knowledge. We believe that the novel knowledge is in different forms in the literature and thus, should fulfil multiple factors to broadly discover them. We observed a similar conclusion from the ARROWSMITH study [11] initiated by the pioneers in LBD discipline and from a recent LBD review [9]. To alleviate the aforementioned limitations, this study attempts to answer the following questions; *does analysing whole time-series in a greater detail benefits in eliciting new knowledge?*, and *does providing a holistic solution that combines the complementary strengths of multiple characteristics (e.g., multiple semantic shifts) yield better predictive effects?*

In summary, our contributions are; 1) proposing a generalisable LBD framework that can be easily adaptable to non-medical LBD settings, 2) quantifying semantic change of topics in conjunction with temporal trajectories and word embedding techniques to capture subtle cues that are robust and highly predictive in suggesting novel knowledge, 3) scrutinising the effect of temporal dynamics with high level of granularity in differentiating the potential connections from the false positives, and 4) integrating machine learning techniques to amalgamate the semantic shift measures to recommend the new knowledge.

2   Related Work

The early work of Swanson demonstrated the potentiality of logically connecting independent information nuggets dispersed across the literature to generate new practical knowledge [9]. Even though these studies formed the groundwork in the discipline, the underlying knowledge synthesis was performed manually requiring a substantial amount of time and effort. Subsequently, several studies [8] have attempted to automate Swanson's manual process by incorporating frequency-based statistical measures. The major limitation of these methods is their excessive dependency on highly frequent topics. Consequently, *RaJoLink* LBD system [8] followed the notion of *rarity* by only favouring the low frequent topics. Nevertheless, reliance on high or rare frequencies were progressive, they do not necessarily capture semantically meaningful connections.

In the meantime, several studies have experimented semantic predications (*subject-predicate-object*) using more specialised medical resources such as

SemRep [6]. Despite being descriptive, the applications of these LBD systems are highly restrictive due to the following reasons. Firstly, they require to have a prior knowledge about all the potential predicates related to the problem and ignore the topics that are outside of the specified predicates. Secondly, the availability of such specialised resources is highly limited to certain problems [8]. Subsequently, another line of research [6] has integrated graph theory to the LBD process by analysing graph properties at; *macro-level* (e.g., shortest path), *meso-level* (e.g., clustering coefficient), and *micro-level* (e.g., centrality measures). Even though the graph-theoretic methods remain more successful, they fail to capture implicit linkages due to their rigid schema [4].

Notwithstanding the research progress gained so far in the discipline, almost all the research studies suffer from the following two major limitations; 1) over-reliance of domain-dependent resources that restricts the model's applications, and 2) neglecting temporal dimension by assuming the literature to be static. Hence, in this study we extend the state-of-the-art techniques by proposing a novel domain-independent temporal methodology to unwind new signals to detect intriguing novel knowledge linkages. Some of the inspiration for this study was emanated from the recent LBD studies that strived to model temporal dimension in LBD process [3,4,12]. Nevertheless, we differ from these studies in multiple ways as discussed in Sect. 1.

3   Overview of the Proposed Model

This section provides a high-level overview of the proposed model by outlining the key functionalities of main phases. Recall that the input to our model is two topics of interest (A and C) and date T, where the goal is to analyse the literature up to time T and to detect latent top k conceptual bridges that are most likely to connect the two topics in future (i.e. in time $T+1$).

To facilitate this, a literature corpus collected up to the time T is required. In this regard, we consider two types of corpora namely; 1) *local corpus:* this is the query specific literature retrieved using the input (i.e. topic A and C) to obtain the local topics, and 2) *global corpus:* this is the entire literature set in the literature database that enables the analysis of local topics in a global scale. Subsequently, the global corpus is split into equivalent sized time-slices to obtain a *time-specific global corpus* that supports evolutionary analysis (see Fig. 2).
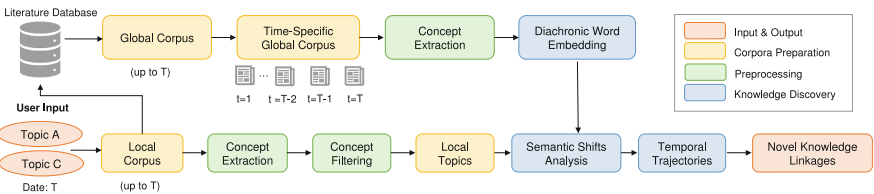

Fig. 2. Overview of the proposed model

In the initial phase of our model, the local corpus is preprocessed (i.e. *concept extraction* and *filtering*) to identify the topics that are local to the input. Subsequently, the evolutionary analysis is performed by considering *semantic change* as the primary temporal setting. To quantify the semantic change, we construct latent embedding spaces for each time window using the global corpus and analyse the temporal trajectories of local topics in a global context. In this regard, we propose three broader classes of evolutionary measures; *individual semantic shifts*, *relative semantic shifts* and *relative semantic shifts extended*.

Next, the derived *semantically infused temporal trajectories* of each local topic are analysed using time-series analysis techniques. To this end, we consider two types of models; *Feature-based Time-series Model (FTM)* and *Dedicated Time-series Model (DTM)*. FTM utilises features extracted from each trajectory that detect patterns in the time-series. Lastly, the features are articulated to recommend the novel knowledge linkages. DTM follows a similar analysis as FTM where we consider the recent advances in deep learning, particularly *Long Short-Term Memory (LSTM)* to learn the patterns from the temporal trajectories. Unlike handcrafted features, such models offer the opportunity in discovering unforeseen structures of novel knowledge.

4   Methodology

4.1   Initial Preprocessing

The first challenge we faced was identifying a suitable *multi-domain* knowledgebase that has a broader coverage than *WordNet* [7] to facilitate the initial preprocessing. In this regard, we selected *DBpedia* which is the *largest multi-domain ontology* lying at the heart of Linked Open Data (LOD) cloud [10] as the primary structured knowledgebase. DBpedia is also a *multilingual* resource which allows this initial preprocessing extendable not only to literature in other domains but also in other languages[1]. To date, DBpedia supports 134 languages. The English version of DBpedia alone includes 1.7 billion facts. We mapped the typical preprocessing steps used in LBD workflow [9] by using the properties of DBpedia as summarised in Table 1. The justification for each DBpedia property selection (in Table 1) and how it is aligned with the LBD workflow are described in Supplementary material Section B. The remaining phases of our methodology do not require any knowledge inferences from outside resources, thereby fulfilling our intention of proposing a *generalisable* LBD model.

4.2   Semantic Shifts

The focus of this section is to discuss how we quantified the semantic change of local topics (*concepts*) to recommend the novel conceptual bridges.

---

[1] Please refer Section A of Supplementary material.

Table 1. Mapping typical preprocessing steps used in LBD using DBpedia

Task	Typical setting	DBpedia setting
Concept identification	Using MeSH or UMLS	DBpedia entities
Synonyms identification	Using synonyms in MeSH or UMLS	*"is dbo:wikiPageRedirects of"* property
Discipline-related terminology	Using controlled vocabulary terminologies such as MeSH, UMLS	"dct:subject" and "skos:broader" properties
Granularity of the concept	Using MeSH hierarchical levels	In-degree page links

The URI prefixes in *DBpedia setting* can be resolved at http://prefix.cc/

Skip Gram with Negative Sampling (SGNS): Since word embeddings can be viewed as a potential diachronic tool [2], we learnt distributed representation of concepts in each distant time-slices of global corpus to analyse how the concepts semantically changed over time. To this end, we utilised the popular neural word embedding; *word2vec* [5] (more specifically *SGNS*) to construct the vector space for each snapshot of global corpus. In these representations each concept w_i has a vector representation $\mathbf{w}^{(t)}$ at each time-slice.

Embedding Alignment: Due to the stochastic nature of SGNS, the constructed word vectors for each time period could be in arbitrary orthogonal transformations. Hence, it is important to align the word vectors to the same co-ordinate axes to facilitate semantic comparison of a same concept across time (e.g., for measures such as global semantic shifts). Defining a matrix of word embeddings trained at time period t as $\mathbf{W}^{(t)} \in \mathbb{R}^{d \times |\mathcal{V}|}$, the orthogonal procrustes alignment was performed across time-periods using Eq. 1 where $\mathbf{R}^{(t)} \in \mathbb{R}^{d \times d}$. The solution corresponds to the best rotational alignment while preserving cosine similarity [2].

$$\mathbf{R}^{(t)} = \underset{\mathbf{Q}^{\top}\mathbf{Q}=\mathbf{I}}{\arg\min} \|\mathbf{Q}\mathbf{W}^{(t)} - \mathbf{W}^{(t+1)}\|^F \tag{1}$$

Measuring Semantic Change: This section outlines how we disentangled multiple types of semantic changes using three broad classes of evolutionary measures to distinguish potential novel knowledge linkages[2].

Individual Semantic Shifts (ISS): In this category, we propose two different ways to measure the semantic shift of an individual concept.

– **Global Semantic Shift:** The global semantic shift quantifies how far a concept has moved in semantic space between two consecutive time periods.

---

[2] Please refer Section C of Supplementary material.

For this purpose, we simply measure the cosine distance of the word vectors of the concept in the aligned vector spaces of time periods t and $t+1$ as in Eq. 2. This measure is sensitive to subtle usage drifts and other global effects [1].

$$d^{\text{ISS-G}}(w_i^{(t)}, w_i^{(t+1)}) = \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_i^{(t+1)}) \qquad (2)$$

– **Local Semantic Shift:** The local semantic shift measures the change of the concept's local neighbourhood. Thus, this measure is sensitive to drastic shifts in core meaning and less sensitive to global shifts. Since, the measure is based on the local semantic neighbours, initially, the concept w_i's $\mathcal{K}$ nearest neighbours at time t are obtained $(\mathcal{N}_{\mathcal{K}}(w_i^{(t)}))$. Subsequently, to quantify the change between the two time-periods t and $t+1$, a second-order similarity vector for $w_i^{(t)}$ is computed from these nearest neighbour sets as defined in Eq. 3. The computed vectors for $w_i^{(t)}$ and $w_i^{(t+1)}$ are used to quantify the local neighbourhood change as in Eq. 4 (see [1] for details).

$$\mathbf{s}^{(t)}(j) = \text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_j^{(t)}) \text{ where } \forall w^j \in \mathcal{N}_{\mathcal{K}}(w_i^{(t)}) \cup \mathcal{N}_{\mathcal{K}}(w_i^{(t+1)}) \qquad (3)$$

$$d^{\text{ISS-L}}(w_i^{(t)}, w_i^{(t+1)}) = \text{cos-dist}(\mathbf{s}_i^{(t)}, \mathbf{s}_i^{(t+1)}) \qquad (4)$$

Relative Semantic Shifts (RSS): In this category, we measure the semantic shifts of the concepts relative to the input topic A and C.

– **Pairwise Semantic Displacement:** This measure quantifies how the semantic similarity of a concept changes over the time relatively to the A and C topics. Thus, this measure verifies if there is a growing semantic similarity of the concept towards topic A and C (see Eq. 5).

$$s^{\text{RSS-S}}(w_i^{(t)}, w_A^{(t)}, w_C^{(t)}) = \text{avg}(\text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}), \text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_C^{(t)})) \qquad (5)$$

– **Pairwise Distance Proximity:** This measure verifies whether a concept's temporal trajectory is leaning towards to a close proximity of both the input topic A and C (Eq. 6). i.e. whether the concept's trajectory is not favouring A or C individually, but both at the same time. The intuition for this measure is that we are seeking for conceptual bridges that implicitly connects A and C, thus, the trajectory should favour both the topics.

$$d^{\text{RSS-D}}(w_i^{(t)}, w_A^{(t)}, w_C^{(t)}) = \max(\text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}), \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_C^{(t)})) \qquad (6)$$
$$+ \beta \mid \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}) - \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}^{(t)}) \mid \text{ where } \beta \geq 0$$

Relative Semantic Shifts Extended (RSSEx): In this category, we extend the two measures proposed in *RSS* category using the recent neighbours of topic A and C namely; **Neighbourhood Semantic Displacement** and **Neighbourhood Semantic Proximity**. The recent neighbours of topic A (N_A) and C (N_C) in a time window $\mathcal{W}$ are calculated as in Eq. 7.

$$N_A = \bigcap_{t=T-\mathcal{W}}^{T} \mathcal{N}_{\mathcal{K}}(w_A^{(t)}), \ N_C = \bigcap_{t=T-\mathcal{W}}^{T} \mathcal{N}_{\mathcal{K}}(w_C^{(t)}) \qquad (7)$$

4.3   Semantically Infused Temporal Trajectories

For each local topic in the corpus, we compute the six semantic shift measures discussed in Sect. 4.2. i.e. every local topic has six temporal trajectories that showcase how the topic semantically changed over the time. The derived semantically infused temporal trajectories are analysed at two levels;

   Feature-based Time-series Model (FTM): This model employs descriptive statistics of each temporal trajectory such as *variance*, *median* as the main temporal features[3]. Considering the variations of semantic shifts, we consider two types of FTM models; 1) *FTM-D:* This type considers *ISS* and *RSS* as the key temporal trajectories, and 2) *FTM-Ex:* This type employs *ISS* and *RSS-Ex* with the intention of evaluating the contribution of local neighbourhood in the relative measures. The potentially of the knowledge linkage is decided by the estimated probability of FTM when the knowledge linkage is in the testing slice.

   Dedicated Time-series Model (DTM): In recent years, *LSTM* models have shown promise in many application areas including time-series and sequential data analysis. Inspired from these research outside of LBD, we employed a sequential LSTM model to analyse the derived temporal trajectories (see footnote 3). Similar to FTM model types, we analyse two types of DTM models namely; *DTM-D* and *DTM-Ex*. The estimated probability of DTM is considered to decide the potentiality of the knowledge linkage when it is in the testing slice.

5   Experiments

Dataset and Test cases Description: We selected MEDLINE as the main dataset of our experiments. It is considered to be the largest scientific repository that provides access to more than 25 million time-stamped scientific articles (mainly in bio-medicine and life sciences) and commonly used as the primary data source of the previous LBD studies [4]. To evaluate the effectiveness of our model and to compare it with the existing models, the following five real-world test cases reported by the pioneers of LBD discipline were selected for re-discovery[4]; 1) Fish-oil (FO) and Raynaud's Disease (RD), 2) Magnesium (MG) and Migraine Disorder (MIG), 3) Somatomedin C (IGF1) and Arginine (ARG),

---

[3] Please refer Section D of Supplementary material.
[4] Please refer Section E of Supplementary material.

4) Alzheimer Disease (AD) Indomethacin (INN), and 5) Schizophrenia (SZ) and Calcium - Independent Phospholipase A2 (PA2).

The main reason for the selection of these test cases is that they are commonly used for LBD evaluation and treated as the *golden datasets* in the discipline. The significance of these test cases in LBD context is that they are *complementary* but *disjoint.* This means the articles in the two topics of each test case have never been mentioned, cited or co-cited each other. Therefore, the aforementioned test cases validate the model's ability in accumulating existing disperse knowledge in literature to develop novel semantic relationships that have *never* drawn any awareness before (i.e.*hypotheses generation*).

Quantitative Evaluation: To evaluate the validity of the generated output, a ground truth formation is required. Unfortunately, LBD discipline does not have any standard ground truth datasets available and constructing such ground truth remains to be an open issue in the discipline. One reason for this is due to the fact that it is unrealistic to build a copious ground truth that will presumably contain all the future discoveries. Hence, the most objective and commonly used quantitative evaluation technique in LBD discipline is to check if the predicted novel discoveries have actually taken place in the future (a.k.a. *time sliced evaluation*) [9]. For this purpose, the literature is divided into two segments using a cut-off-date namely; 1) *pre-cut-off segment:* where the literature in this segment are used as the input to the LBD models to discover the novel knowledge and 2) *post cut-off segment:* where the literature in this segment are used to verify if the predicted discoveries of the LBD models have actually been discovered and published. In other words, a generated discovery is considered to be valid only if it is absent in pre cut-off segment (i.e. the predicted discovery has not taken place until the cut-off-date) and present in the post cut-off segment (i.e. the predicted discovery has been discovered and published after the cut-off-date) (see footnote 4).

Evaluation Setting: In the evaluation setup we validate the coverage of ground truth conceptual bridges in the top k recommendations. In other words, it is vital to prioritise the detected conceptual bridges in a way where the topmost recommendations represent accurate novel knowledge. For this purpose, similar to previous LBD studies, [3] the two evaluation metrics; *Precision@k (P@k)* and *Mean Average Precision@k (MAP@k)* are used to quantify the results.

Evaluation Baselines: Same as prior LBD studies [3,4,12], the following *eight* baseline algorithms were used for performance comparison; 1) *Dynamic Embeddings (DE)*[5], 2) *ARROWSMITH (AR)*, 3) *Mutual Information (MM)*, 4) *Apriori algorithm (AP)*, 5) *TF-IDF (TI):*, 6) *Literature Cohesiveness (COH)*, 7) *Static Embeddings (SE)*, and 8) *Static Networks (SN)*.

---

[5] How our model differs from this recent work [12] is discussed in Sect. 1.

5.1   Results and Discussion

Table 2 reports the *P@k* for the golden datasets; (1) and (2) where *k* is progressively increased from 10 to 100. The *P@k* result tables for all the five golden test cases are reported in Section F of Supplementary material. While *P@k* indicates the coverage of correct recommendations, *MAP@k* (which is the arithmetic mean of *Average Precision@k*) measures the overall performance of the models considering their ranking order of correct recommendations. Table 3 reports the *MAP@k* of the five golden test cases. The results of both *P@k* and *MAP@k* indicate that all the variants of the proposed model consistently outperform the existing baselines which demonstrates the ability of detailed semantic shifts analysis in detecting meaningful novel knowledge linkages.

We revealed the following trends through the analysis of the proposed variants of the our model. In terms of the *coverage* of correct recommendations (i.e. *P@k*), *DTM-D* reports the highest overall performance across datasets. This showcase the ability of LSTMs in detecting unforeseen structures in the temporal trajectories that are useful in differentiating potential new knowledge from false

Table 2. Precision@k for FO-RD and MIG-MG golden test cases

Methods	10	20	30	40	50	60	70	80	90	100
Fish-oil and Raynaud's Disease										
DE	0.2	0.2	0.2	0.2	0.2	0.25	0.2286	0.225	0.2222	0.22
AR	0.4	0.5	0.5333	0.475	0.46	0.4667	0.5	0.475	0.4778	0.51
MM	0.4	0.25	0.2667	0.25	0.28	0.3333	0.3286	0.3625	0.3889	0.38
AP	0.1	0.15	0.3333	0.35	0.38	0.4	0.3571	0.35	0.3444	0.35
TI	0.6	0.45	0.3	0.275	0.26	0.2333	0.2714	0.275	0.2778	0.27
COH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01
SE	0.0	0.15	0.2333	0.2	0.2	0.2667	0.2714	0.2875	0.2889	0.32
SN	0.0	0.15	0.2	0.175	0.24	0.2833	0.3	0.3125	0.3222	0.34
FTM-D	0.7	0.8	0.8333	0.875	0.84	0.8167	0.7714	0.7625	0.7556	0.76
FTM-Ex	0.9	**0.95**	**0.9333**	**0.925**	**0.9**	0.8167	0.8	0.8	0.8111	0.81
DTM-D	**1.0**	**0.95**	0.9	0.9	**0.9**	**0.9**	**0.9**	**0.8875**	**0.8778**	**0.88**
DTM-Ex	0.9	0.85	0.8667	0.825	0.86	0.8167	0.8143	0.8	0.8111	0.81
Magnesium and Migraine Disorder										
DE	0.2	0.2	0.1667	0.2	0.2	0.2	0.2143	0.1875	0.1889	0.21
AR	0.6	0.5	0.5333	0.525	0.54	0.5167	0.5143	0.55	0.5444	0.58
MM	0.2	0.4	0.4667	0.475	0.42	0.4	0.3571	0.3375	0.3222	0.33
AP	0.0	0.1	0.1333	0125	0.12	0.15	0.1714	0.175	0.2	0.22
TI	0.2	0.4	0.5	0.475	0.46	0.45	0.4286	0.4125	0.3778	0.36
COH	0.0	0.0	0.0	0.0	0.0	0.0167	0.0286	0.0375	0.0556	0.05
SE	0.4	0.35	0.2333	0.25	0.24	0.2667	0.2714	0.3	0.2889	0.28
SN	0.0	0.0	0.0333	0.05	0.04	0.05	0.0429	0.0375	0.0667	0.07
FTM-D	0.5	0.65	0.7	0.775	0.76	0.7833	0.8	**0.8**	**0.7889**	**0.8**
FTM-Ex	0.6	0.7	0.7333	**0.8**	**0.82**	**0.8333**	**0.8429**	**0.8**	0.7667	0.74
DTM-D	**0.8**	**0.85**	**0.8333**	0.75	0.74	0.7667	0.7571	0.775	**0.7889**	**0.8**
DTM-Ex	0.7	0.75	0.7333	0.775	0.78	0.7833	0.7714	**0.8**	**0.7889**	0.77

Table 3. Mean Average Precision@k (MAP@k) for all golden test cases

Methods	10	20	30	40	50	60	70	80	90	100
DE	0.0768	0.0647	0.0648	0.0599	0.0534	0.0561	0.0529	0.0480	0.0466	0.0461
AR	0.2701	0.2722	0.2738	0.2573	0.2520	0.2511	0.2552	0.2574	0.2558	0.2703
MM	0.1526	0.1510	0.1559	0.1546	0.1622	0.1786	0.1723	0.1726	0.1746	0.1710
AP	0.0480	0.0407	0.0569	0.0606	0.0609	0.0610	0.0576	0.0570	0.0573	0.0612
TI	0.1593	0.1329	0.1098	0.0973	0.0877	0.0810	0.0777	0.0747	0.0697	0.0662
COH	0.0	0.0	0.0	0.0	0.0	0.0	0.0001	0.0002	0.0005	0.0004
SE	0.0433	0.0462	0.0510	0.0471	0.0491	0.0613	0.0652	0.0718	0.0742	0.0805
SN	0.02	0.0194	0.0240	0.0209	0.0253	0.0292	0.0294	0.0331	0.0347	0.0341
FTM-D	0.4630	0.5671	0.5742	0.5951	0.5830	0.5762	0.5591	0.5515	0.5420	0.5348
FTM-Ex	**0.6283**	**0.6813**	**0.6639**	**0.6895**	**0.6760**	**0.6391**	**0.6303**	**0.6067**	**0.6004**	**0.5943**
DTM-D	0.6096	0.5596	0.5407	0.5135	0.5174	0.5297	0.5272	0.5288	0.5391	0.5455
DTM-Ex	0.6054	0.6018	0.5647	0.5384	0.5360	0.5073	0.5094	0.5028	0.4985	0.5001

connections. The *MAP@k* results indicate that *FTM-Ex* consistently outperform the remaining models by often front-loading the correct recommendations. i.e. this model tend to have a *better ordering* of the knowledge recommendations. This highlights that the LBD model is sensitive not only to the topic A and C alone, but also to their core meaning. In recommendation tasks such as LBD, it is unrealistic to expect that the user will examine and experiment the entire list of proposed new knowledge linkages. In other words, *better ordering* of the knowledge recommendations is crucial compared to *coverage*. Therefore, we believe that the slight *P@k* loss incurred using *FTM-Ex* in most of the test cases can be indemnified through its performance gain in *MAP@k*.

Overall, we consider the following reasons as the main strengths of the proposed model compared to the baselines; 1) *Multiple characteristics:* Even though most of the LBD studies strictly rely on one or two characteristics to elicit new knowledge, it is observed that new knowledge can be in multiple forms due to complexity of the knowledge structures in the scientific literature. For instance, Davies [9] has identified five forms of novel knowledge in FO-RD and MG-MIG test cases. Thus, our model take the advantage of detecting novel knowledge in different forms by capturing the semantic change at multiple levels; *individual shifts*, *pairwise shifts* and *neighbourhood shifts*, 2) *Global analysis:* Unlike most of the prior LBD research that merely focus on cues at local scale, we analysed the concepts' trajectories in a global context. This facilitates the analysis of concepts neighbourhood in a wider scope. For example, consider "blood viscosity" conceptual bridge of FO-RD test case. This conceptual bridge may also be associated with other chemical substances of FO (such as *eicosapentaenoic acid*). However, query specific local corpus often limits in accommodating such implicit interactions, 3) *Detailed temporal analysis:* While almost all the prior LBD research are based on the static literature, we considered the temporal behaviour of concepts to discover new knowledge. This allows the model to detect time-aware knowledge recommendations that have higher semantic meaning. Moreover, it is also evident that analysing the time-series in detail benefits in LBD workflow (in contrast to baselines such as *DE* [12]), and 4) *Generalisability:* The proposed temporal clues are free from knowledge

inferences from domain-dependent resources (unlike baselines such as AR [11]). This meets our objective of generalisabe cues whose predictive effects do not rely on the specialised knowledgebases. Moreover, our initial preprocessing phase is also adaptable to multiple domains and languages due to strengths of DBpedia. Thus, our solution can be easily integrated to non-medical LBD settings.

Further analysing the results, we observe that AR performs the best among the baselines. AR is arguably the most popular and well-maintained LBD system in the discipline that currently has nearly 1200 of monthly user-base [6]. We observe two main reasons for its performance gain compared to the remaining baselines. Firstly, it considers seven characteristics to determine the potentiality of the novel knowledge (i.e. the use of *multiple characteristics*). Secondly, three of their characteristics include global literature analysis which benefits in identifying the concept's *global properties* that are not visible to local corpus. However, three of its features require the analysis of *UMLS* and *MeSH*, which restricts the suitability of AR baseline only to the medical domain.

6   Conclusion and Future Work

In this study, we have described, evaluated and systematically compared our semantically infused temporal model in detecting novel knowledge linkages. The results indicate the challenge associated in detecting such novel linkages and emphasis the need of developing circumstantial solutions to handle the problem. Overall, the holistic integration of semantics and temporal information significantly outperformed all the existing baselines in the discipline. The supplementary material of this paper is also available at: https://tinyurl.com/lbd-supplementary.

In future research, we intend to take the advantage of the power of the proposed semantic shifts and the domain-independency of the model to contribute to LBD research in non-medical domains such as *computer science* (thus far, there exists only one LBD study in *computer science* [8]). Therefore, we believe that our model will be a successful first step towards promoting generalisable LBD systems.

References

1. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Cultural shift or linguistic drift? comparing two computational measures of semantic change. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2016, p. 2116. NIH Public Access (2016)
2. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1489–1501 (2016)
3. Jha, K., Xun, G., Wang, Y., Gopalakrishnan, V., Zhang, A.: Concepts-bridges: uncovering conceptual bridges based on biomedical concept evolution. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1599–1607. ACM (2018)

4. Jha, K., Xun, G., Wang, Y., Zhang, A.: Hypothesis generation from text based on co-evolution of biomedical concepts. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 843–851. ACM (2019)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
6. Sebastian, Y., Siew, E.G., Orimaye, S.O.: Emerging approaches in literature-based discovery: techniques and performance review. Knowl. Eng. Rev. **32**, 1–35 (2017)
7. Sebastian, Y., Siew, E.G., Orimaye, S.O.: Learning the heterogeneous bibliographic information network for literature-based discovery. Knowl.-Based Syst. **115**, 66–79 (2017)
8. Thilakaratne, M., Falkner, K., Atapattu, T.: A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. ACM Comput. Surv. **52**(6), 1–34 (2019)
9. Thilakaratne, M., Falkner, K., Atapattu, T.: A systematic review on literature-based discovery workflow. PeerJ Comput. Sci. **5**, e235 (2019)
10. Titze, G., Bryl, V., Zirn, C., Ponzetto, S.P.: DBpedia domains: augmenting DBpedia with domain information. In: LREC, pp. 1438–1442 (2014)
11. Torvik, V.I., Smalheiser, N.R.: A quantitative model for linking two disparate sets of articles in medline. Bioinformatics **23**(13), 1658–1665 (2007)
12. Xun, G., Jha, K., Gopalakrishnan, V., Li, Y., Zhang, A.: Generating medical hypotheses based on evolutionary medical concepts. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 535–544. IEEE (2017)