Visual Informatics 5 (2021) 1–12

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

USEVis: Visual analytics of attention-based neural embedding in information retrieval

Xiaonan Ji [a,b,*], Yamei Tu [b], Wenbin He [b], Junpeng Wang [b], Han-Wei Shen [b], Po-Yin Yen [a]

[a] Institute for Informatics, Washington University School of Medicine in St. Louis, United States of America
[b] Computer Science and Engineering, The Ohio State University, United States of America

ARTICLE INFO

Article history:
Available online 2 April 2021

Keywords:
Interactive visual system
Neural embedding
Attention mechanism
Document understanding
Information retrieval
Clinical decision-making

ABSTRACT

Neural attention-based encoders, which effectively attend sentence tokens to their associated context without being restricted by long-term distance or dependency, have demonstrated outstanding performance in embedding sentences into meaningful representations (embeddings). The Universal Sentence Encoder (USE) is one of the most well-recognized deep neural network (DNN) based solutions, which is facilitated with an attention-driven transformer architecture and has been pre-trained on a large number of sentences from the Internet. Besides the fact that USE has been widely used in many downstream applications, including information retrieval (IR), interpreting its complicated internal working mechanism remains challenging. In this work, we present a visual analytics solution towards addressing this challenge. Specifically, focused on semantics and syntactics (concepts and relations) that are critical to domain clinical IR, we designed and developed a visual analytics system, i.e., USEVis. The system investigates the power of USE in effectively extracting sentences' semantics and syntactics through exploring and interpreting how linguistic properties are captured by attentions. Furthermore, by thoroughly examining and comparing the inherent patterns of these attentions, we are able to exploit attentions to retrieve sentences/documents that have similar semantics or are closely related to a given clinical problem in IR. By collaborating with domain experts, we demonstrate use cases with inspiring findings to validate the contribution of our work and the effectiveness of our system.

1. Introduction

Neural embedding converts unstructured texts (e.g., words, sentences, or documents) to structured vector representations that encode the underlying texts' meanings (Lopez and Kalita, 2017). The vector representations, i.e., embeddings, can then be used to improve the effectiveness and efficiency in many downstream machine learning tasks, e.g., text classification or clustering. Information retrieval (IR) (Mitra and Craswell, 2017) is one of the multiple applications that has been significantly benefited from neural embedding.

Recently, neural network models with the attention mechanism, e.g., transformer, have demonstrated advantages in producing high-quality embeddings (Bahdanau et al., 2014; Vaswani et al., 2017). As one of the most influential topics in deep learning and neural embedding, the attention mechanism associates related tokens (words) across texts without being restricted by long-term dependency (i.e., taking the entire context into account). Some of these associated words further constitute linguistic properties (such as semantic concepts and syntactic relations) accounting for underlying text meanings, thus promoting the quality of resulting embeddings.

As the attention mechanism has been widely adopted in producing text embeddings for various applications, interpreting attentions plays an important role in understanding the merits as well as limitations of attentions-based models. Moreover, for applications that have strong safety concerns, such as biomedical and clinical IR, a thorough understanding of the attention mechanism is crucial for either the safety of individual patients or the management of the global health. However, due to the highly complex architectures of attention-based models (e.g., multi-head multi-layer attentions, Cer et al., 2018; Devlin et al., 2018), interpreting and analyzing attentions are challenging. Previous works Vaswani et al. (2017) and Vig (2019b,a) have leveraged the power of visualization to interpret attentions at the word level in sample sentences, which is inspiring but has limitations in analyzing attentions at the sentence (or document) level or exploiting attentions for IR.

In this study, we propose a visual analytics approach to gain insights into attention patterns across all sentence words and

* Corresponding author at: Institute for Informatics, Washington University School of Medicine in St. Louis, United States of America.
E-mail address: ji.62@osu.edu (X. Ji).

https://doi.org/10.1016/j.visinf.2021.03.003

probe the linguistic properties captured by attentions. More specifically, we focus on IR applications (such as document classification, clustering, recommendation, etc.) that strongly rely on effective feature representations, and can benefit from attention-based neural embedding, as well as insights into the attention mechanism. With such an application-driven view, we aim to contribute to clinical IR that are critical to real-world healthcare decision-making. Under this notion, we are specially interested in how domain concepts and relations, e.g., semantics or syntactics related to patient, disease, treatment, outcome, are captured by attentions. Furthermore, with iterative discussions with domain experts, we are also interested in exploiting interpretable and comparable attention patterns to assist in IR applications with human-in-the-loop. Therefore, our objectives include (1) interpreting attentions in an IR guided manner, and (2) exploiting the interpretable attentions to support IR applications. These two objectives are mutually supportive to each other.

With the aforementioned objectives and rationales, we developed an interactive visual analytics tool to interpret, explore, and exploit attention patterns from multiple levels: instance-level (an individual sentence), group-level (a small set of related sentences), document-level (a document consisting of multiple sentences), and corpus-level (an IR corpus of documents). With document-level and corpus-level analytics, this study is also designed with respect to real-world IR applications towards a document corpus. In summary, our work contributes to:

- Explore linguistic properties (semantics or syntactics) encoded in attentions with a visual analytics approach.
- Explore the roles of attention heads in capturing specific linguistic properties. This is with respect to the multi-head nature of widely used attention models, such as USE (Cer et al., 2018) and BERT (Devlin et al., 2018).
- Exploit interpretable and comparable attention patterns to assist in IR applications. This is with respect to requirements from practical IR, as well as multi-faceted text being used. Attention patterns could suggest salient information and essential meanings that are of task interest.

2. Related work

In this section, we review related work in visual analytics of neural embedding and bi-directional relation visualization.

2.1. Visual analytics of neural embedding

Most of the previous work of visualizing neural embedding models focused on deep recurrent models, e.g., Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). Kiros et al. (2015) evaluated sentence encodings of GRU through visual analytics. Smilkov et al. (2016) proposed the Embedding Projector, a visual analytics system to interpret neural embeddings and evaluate their performances. Palangi et al. (2016) visualized the activation behaviors of LSTM models with respect to various sentences. Lopez and Kalita (2017) visualized document representations generated with a dynamic convolutional neural network. Liu et al. (2018) focused on visualizing and analyzing the relations between word embeddings. Ming et al. (2017) visualized the relation between embeddings and sentences with biclustering for RNN models. Strobelt et al. (2018) proposed a visual analytics system to help experts form and verify hypotheses on neural embeddings generated by LSTM models. In our previous study (Ji et al., 2019), we conducted visual exploration of neural document embedding, which was based on the well-known Paragraph Vector model (i.e., doc2vec, Le and Mikolov, 2014) but is generalizable to embeddings produced by other neural models.

Recently, attention-based models have shown outstanding performance in solving NLP tasks, especially for embedding sentences into meaningful representations. A few pioneering works have been proposed to visualize and analyze attention-based models. Vaswani et al. (2017) proposed the attention-driven transformer architecture, and visualized attentions pertained to individual words in a sample sentence. Lin et al. (2019) visualized attentions by annotating on the sentences. Vig (2019b,a) visualized attentions within a sentence by connecting the words with links weighted by the attentions. Park et al. (2019) proposed SANVis, a visual analytics tool to understand the attention mechanism of transformer in NLP scenarios.

Our work is highly inspired by these pioneering studies in visual analytics of neural embeddings. While existing techniques mostly focused on understanding embeddings' performance, behavior, or linguistic properties, our work aims to understand the underlying mechanism of high-quality embeddings, with respect to an attention-driven model. Thus the objective of visual design is different, for instance, LSTMVis (Strobelt et al., 2018) was towards the hidden state dynamics, while our work is towards the attention mechanism. Furthermore, comparing to existing techniques focusing on the attention mechanism (especially self-attention with transformer), our work takes an IR application-drive view, aiming to not only understand but also exploit attentions for IR. Thus the scope and implementation of visual analytics are different, for instance, SANVis (Park et al., 2019) interpreted attention patterns across multiple heads, while our work interprets attentions related to semantic concepts or syntactic relations across multiple heads, and promotes interpretable attentions to the document-level. In summary, being inspired by the pioneering studies, our work brings novel contributions in visualizing attention patterns across a sentence or even a document (consisting of multiple sentences), and promoting intuitive interpretation and interactive exploration with human in-the-loop for IR.

2.2. Bi-directional relation visualization

Attentions are essentially the encoding of the bi-directional relations between sentence tokens or words. Various techniques have been proposed to visualize bi-directional relations, among which the node-link diagram and adjacency matrix are the most popular approaches. Node-link diagrams visualize the relations as a graph, where each link between a pair of nodes represents the relation between the nodes. For bi-directional relations, arrow links are often used. The key challenge of this group of visualization approaches is how to layout the graph, for which various layout methods have been proposed, such as force directed layouts (Battista et al., 1994), spectral layout (Koren, 2005), tree layout (Herman et al., 2000), etc. Adjacency matrices are also frequently used to visualize bi-directional relations, where each cell in the matrix represents the relation between two objects, and each row/column represents the relations between one object to all the others. Various techniques have been proposed to improve the usability and readability of the matrix-based approach, such as MatrixExplorer (Henry and Fekete, 2006), MatLink (Henry Riche and Fekete, 2007), and NodeTrix (Henry et al., 2007). Among pioneering studies, the visual analytics system of Jigsaw (Görg et al., 2014) employed a variety of visualization techniques (e.g., list, graph, connection-based views) to explore bi-directional relations among entities extracted from document collections; and BiDots (Zhao et al., 2017) developed interactive bi-clustering for document analysis tasks. In this work towards sentences and documents, we visualize attentions with both node-link diagram (*graph-based visualization*) and adjacency matrix (*heatmap-based visualization*) to employ the advantage of both methods and conduct more comprehensive visual analytical tasks.

3. Background on attention mechanism

Attention has been playing an increasingly important role in deep learning for various applications, such as reading comprehension, text entailment, abstractive summarization, image captioning (Cui et al., 2016; Chen and Zhuge, 2018). In this work, we focus on visualizing and analyzing attention models applied on text mining and IR, which the attention mechanism was originally designed for. In the following, we first discuss the original attention model used to facilitate sequential models for neural machine translation. Then, we move on to the self-attention and universal sentence encoder, which are the main focus of this work.

3.1. Attention and neural sequence modeling

Attention was originally designed for neural machine translation and sequence modeling with a seq2seq architecture (Bahdanau et al., 2014). A seq2seq architecture generally includes an encoder and a decoder (Fig. 1), both of which are usually realized with RNN (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014). In a machine translation problem, the encoder sequentially reads in a sentence $(x_1, \ldots, x_T)$ from the source language, while updating and utilizing its hidden states $(h_1, \ldots, h_T)$. Therefore, with the recurrent units, the source sentence's context information is accumulated into the encoder's final states h_T, which are further fed into the decoder as its initial states s_0. The decoder then sequentially produces the target sentence $(y_1, \ldots, y_m)$ while updating its hidden states and consuming the previous prediction.

With an attention mechanism, the prediction of a target word y_i can attend to the encoder's every intermediate state (e.g., h_j), corresponding to every source word (e.g., x_j). Moreover, attention weights are learned to suggest how much attention the decoder should pay to h_j when predicting y_i. Specifically, for the target word y_i, the attention weight for a source word x_j can be noted as $\alpha(i, j)$, which is based on a softmax normalization of $e(i, j)$ that is calculated as the compatibility between s_{i-1} and h_j via an alignment model, i.e., $e(i, j) = a(s_{i-1}, h_j)$. Consequently, a context vector is generated for y_i using a weighted sum of the encoder's intermediate states $(h_i, \ldots, h_T)$. With the context vector being utilized to predict the target word, the source context is leveraged in a more comprehensive manner.

3.2. Self-attention and universal sentence encoder

The transformer (Vaswani et al., 2017) architecture with self-attention (intra-attention) is a continuing effort and the current state-of-the-art process. It follows the encoder–decoder structure, with stacked self-attention (i.e., 6 self-attention layers) used in both encoder and decoder, as illustrated in Fig. 2. Entirely reliant on self-attention to compute representations, the transformer does not involve recurrent or convolutional units. With self-attention, different words (tokens) within a sentence are related to each other in order to produce a representation of the sentence. Therefore, self-attention is a successful extension to the classic attention mechanism as described in Section 3.1. Instead of attending words across sentences, self-attention has a within-sentence scope and the advantage of capturing intrinsic sentence patterns, such as linguistic properties constituted by sentence words. In fact, the encoding sub-graph of the transformer architecture is extracted and used as a sentence encoder in neural embedding, and it is known as Universal Sentence Encoder (USE) (Cer et al., 2018).

Basically, with self-attention, each sentence word w_i comes with a triplet (q_i, k_i, v_i), each component is an intermediate tensor value (e.g., *512-dimensional vector*) representing the *query*, *key*, or *value* learned for w_i. To compute the attention weight from w_i (source word) to w_j (target word), which suggests their compatibility or association, a dot product between q_i and k_j is used to compute $\alpha(i, j)$. For w_i's context vector, which is a context-aware representation, a weighted sum of all (target) words' value vectors is computed. Importantly, the attention weight is directed, for example, $\alpha(i, j)$ indicates how much attention w_i pays to w_j or how much attention w_j receives from w_i. A word w_i's context vector is representing how w_i pays attentions to other words across the sentence. The attentions would associate related words and account for linguistic patterns that are important for the underlying sentence meanings.

With USE, each of the 6 self-attention layers (stacked and sequential) performs the aforementioned operations and generates context vectors in a gradually refined manner, e.g., from lower-level patterns to high-level patterns. Furthermore, each attention layer contains 8 attention heads in parallel. Such multi-head attention allows the model to jointly attend to sentence context from different perspectives or different representation sub-spaces. In fact, USE's final output, i.e., the resulting sentence embedding, highly relies on the attentions and the corresponding context vectors in the intermediate attention layers. Therefore, interpreting attentions would help explain USE's demonstrated superior performance and provide insights into potential limitations and improvements. Furthermore, as a general-purpose and transferable model, USE is pre-trained by extensive online resources (e.g., Wikipedia, web news, web question–answer, Stanford Natural Language Inference corpus) with both unsupervised and supervised training schemes. It is also of interest to investigate USE's adaptation into domain applications, i.e., whether the attentions can capture semantic or syntactic patterns meaningful to the domain.

In this study, we approached the pre-trained USE model from TensorFlow Hub (https://tfhub.dev/google/universal-sentence-encoder/), fed in a sentence, and extracted intermediate tensor values corresponding to the learned attention weights. We also referred to the list of tokens as processed by USE: all sentence words are preserved (with the removal of punctuations) and two special tokens are added, i.e., ⟨s⟩ for sentence start, and ⟨/s⟩ for sentence end.

4. Design process

4.1. Goals and interests

With an extensibility to BERT and other Transformer based models, we aim to interpret and explore linguistic properties (semantics and syntactics) encoded in USE's multi-layer and multi-head attentions, which enable the generation of high-quality sentence embeddings benefiting downstream applications. We also aim to leverage interpretable attentions to assist in real-world IR with human-in-the-loop.

With an application-driven view, we collaborated with domain experts who are experienced in text mining and IR. We specifically focus on clinical IR, which requires identifications of relevant documents (e.g., clinical trials) to inform healthcare decision-making for critical patient problems. The IR performance highly relies on feature representations (embeddings) that can effectively encode underlying text meanings. Among the multi-faceted text meanings, the most important aspects for clinical IR typically include: clinical study design, patient population, disease, treatment, outcomes, etc. Therefore, besides applying USE to generate sentence/document embeddings for clinical IR, we wanted to probe how the attention mechanism encodes semantic and syntactic patterns related to the afore-mentioned

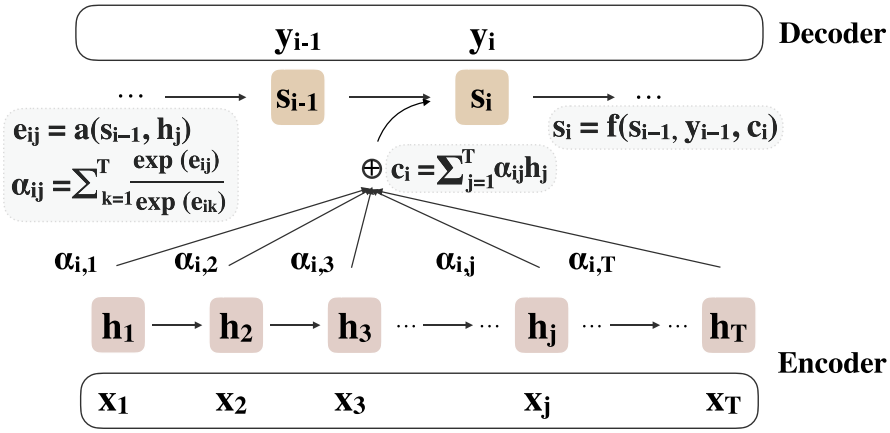

Fig. 1. Seq2seq model for machine translation, which uses attention mechanism to capture long-term dependencies.

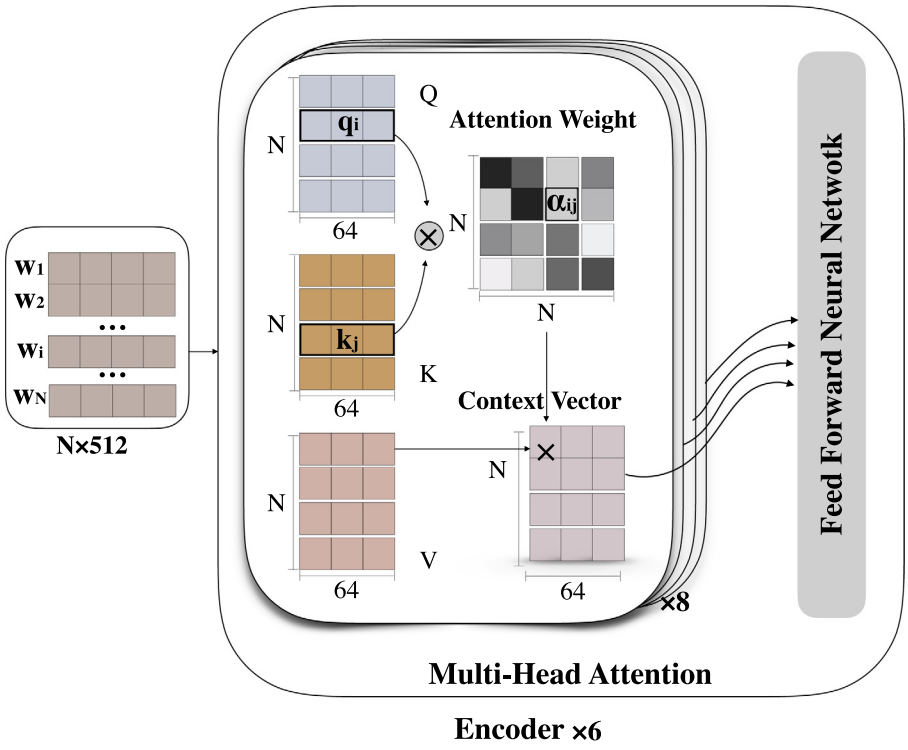

Fig. 2. Universal sentence encoder with stacked self-attention layers.

aspects. For example, how attentions associate multiple words to form a semantic concept of a disease name (e.g., *attention deficit hyperactivity disorder*), how attentions are derived from words indicating a treatment to words indicating the targeting disease, etc. Under this notion, interpretable attentions can be beneficial to downstream IR, lies in that (1) gaining confidence in the generated embeddings, (2) providing a venue to understand documents with revealed semantics and syntactics, (3) facilitating to retrieve desired documents through the revealed semantics and syntactics with human-in-the-loop enabled (thus steering with domain knowledge or task interest).

4.2. Domain inputs

We had an iterative process to design and develop USEVis to meet our goals and interests. With domain experts involved in our research group, we held weekly or biweekly meetings to distill visual analytical tasks, prototype visual design, discuss analytical results, and obtain iterative domain feedback. Below we highlight some of the inspiring feedback and beneficial refinement from discussions with domain experts:

Focused Concepts and Relations. While there are a variety of linguistic properties that we can probe with the attention mechanism, it is more practical and meaningful to focus on properties of domain IR interest. As elaborated by domain experts, these properties include semantic concepts indicating patient population, disease, treatment, outcome, etc.; semantic or syntactic relations about patient-and-disease, disease-and-treatment, treatment-and-outcome, etc.; and other helpful syntactic patterns reflecting sentence structure/type, e.g., a sentence *questioning* about the outcome of a treatment.

Multi-head Attentions. By analyzing experimental results with domain experts, we found different attention heads of USE tend to capture different linguistic properties. For example, one attention head might direct attentions to a few keywords, another

might emphasize functional words for the sentence structure. This inspired us to probe potential roles pertained to certain attention heads. Therefore, our interests are enriched to (a) understand the mechanism of multi-head attentions, and (b) investigate whether the multi-head design can account for the multi-faceted nature of text meanings.

Document-level Attentions. USE is mainly developed for and applied to sentence-level context, thus attention patterns are usually analyzed with a sentence-level granularity, which is considered a coherent context. Furthermore, domain experts pointed out that many IR applications are conducted towards a document corpus, and suggested the necessity to explore attention patterns with a document-level granularity. Thus we proposed to treat a document as a batch of sentences, and synthesize attentions from sentences to a document.

Utilizing Attentions in IR. Our interest of exploiting interpretable attentions in IR is further motivated by the multi-faceted nature of sentences or documents. For example, a sentence can encapsulate information about a disease name and the treatment outcome; similarly, a clinical trial can enclose even more diverse information, e.g., ranging from clinical study design, patient symptom, treatment procedure, to result evaluation. Domain experts commented that interpretable attentions could provide an avenue to recognize a particular facet of interest, thus retrieve sentences or documents delivering the desired facet. The multi-head attentions could be of help for this purpose.

Multi-level Analysis. We will analyze attention patterns in four levels: (a) Instance-level for one individual sentence. This allows us to gain detailed understanding of semantics and syntactics encoded by an attention head in a sentence context. (b) Group-level for a group of sentences with similar syntactic, semantics, or topics. This enables examinations of one attention head's potential role across different (but controlled) sentences. (c) Document-level for multiple sentences belonging to one document. This synthesizes attention patterns for a document. (d) Corpus-level for an IR corpus of documents, where we look for similar attention patterns across documents. While (a)(b) gain insights and confidence in USE's attention mechanism, (c)(d) prepare for IR applications towards a document or a corpus.

4.3. Analytical tasks

Based on the iterative design sessions, below we distill a list of visual analytical tasks, T1–T4, to guide the development of USEVis.

T1: (instance-level task) Interpret what linguistic properties, i.e., semantics or syntactics, are encoded in attentions. We focus on domain concepts and relations used in clinical IR. For a coherent context, we focus on an individual sentence and examine bi-directional relations across sentence tokens, as captured by any attention head. This task is aligned with the domain insight in Focused Concepts and Relations.

T2: (group-level task) Probe linguistic properties pertained to attention heads. For each attention head, we probe whether consistent properties are captured across different sentences; Given multi-head attentions, we probe whether different roles are performed. This task is inspired by the domain insight in Multi-head Attentions.

T3: (document-level task) Explore attentions for a document consisting of multiple sentences. As an extension to sentence-level attentions, we address domain concerns and synthesize attentions from multiple sentences for their belonged document. We also explore the potential to leverage attentions for document understanding. This task is motivated by the domain insight in Document-level Attentions.

T4: (corpus-level task) Exploit interpretable attentions to assist in IR. For domain IR towards a corpus of documents, we explore the potential to facilitate IR by retrieving documents based on recognized attention patterns of interest. This aligns with the multi-faceted nature of documents, and leverages the multi-head attention patterns by USE. This task is motivated by the domain insight in Utilizing Attentions in IR.

Motivated by the domain insight in Multi-level Analysis, these tasks form a bottom-up process and are related to each other as follows. T1 interprets attentions of any head for an individual sentence, with an elementary but essential granularity. T2 then approaches the multi-head nature and probes pertained properties or roles, via summarizing attentions of any head across a group of sentences. T1 and T2 are fundamental tasks to interpret attentions with respect to a multi-head mechanism. Furthermore, T3 and T4, which are built upon T1 and T2, promote attentions to IR tasks. T3 synthesizes attentions for a document consisting of multiple sentences. T4 then exploits interpretable attentions for IR towards a corpus of documents. For T3 and T4, the multi-faceted document meanings could be approached by the multi-head attentions.

5. Visual analytics system

USEVis is mainly composed of four components as follows, with corresponding analytical tasks indicated in brackets. To better illustrate the rationale of these building blocks, we also make correspondences to domain questions in IR.

(1) Visualize attention patterns for a sentence with respect to a selected attention head (from a selected attention layer) [T1, T2]. It enables the visualization and exploratory interpretation of attention patterns at the fundamental sentence-level (instance-level), and helps probe the encoded semantics and syntactics. It addresses the question *what domain concepts or relations are encoded by an attention head?* Besides, by switching among different attention heads, it also helps with *what are the multiple facets of one sentence being captured by USE's multiple attention heads?*

(2) Summarize attention patterns across a group of sentences for a selected attention head; and enable comparisons across multiple attention heads [T2]. It accumulates information from multiple sentences and visually presents the most salient properties captured by an attention head, for an intuitive overview and cross-head comparisons. Therefore, it addresses questions *does an attention head tend to have a specific role and capture specific linguistic properties even for different sentences?* and *do different attention heads contribute to different linguistic properties?*

(3) Synthesize attention patterns for a document consisting of multiple sentences [T3, T4]. As sentences are building blocks of a document, sentence semantics composes document semantics accordingly. By putting together sentence-level attentions, this component illustrates document-level attentions with a synthetic visualization. Besides, it also provides a possible avenue for *document understanding* and helps with the domain question to *recognize relevant documents of domain or task interest in IR.*

(4) Generate representations (e.g., vector representations or embeddings) of attention patterns and support the retrieval of similar attention patterns [T4]. With a recognized document which encloses an interpretable attention pattern indicating relevant semantics, a followup domain question raised in IR is to *retrieve additional documents enclosing similar semantics while mitigating noises from other facets*. This component generates a vector representation for each attention pattern, so that we can have automatic comparisons across different attention patterns and identify similar ones.

We put together these components for **utilizing interpretable and comparable attentions to support IR towards a document**

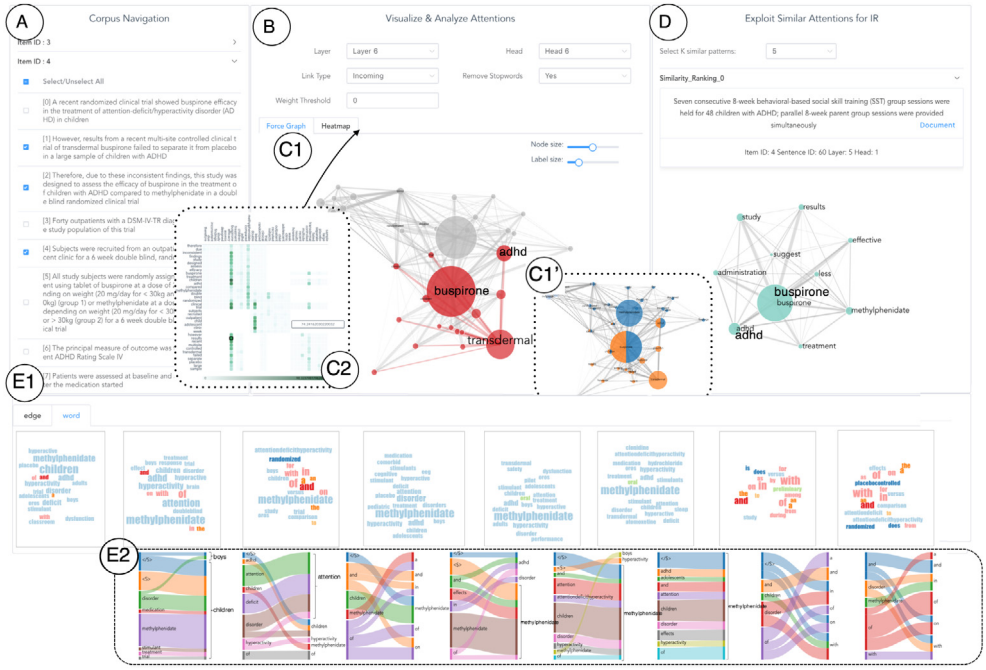

Fig. 3. (A) Corpus navigation, users can select a document and a set of enclosed sentences to explore attention patterns encoded by USE. We use a clinical corpus from domain information retrieval (IR). (B) Attention visualization control, users can specify an attention layer and head, and filter sentence words (nodes) and relations/attentions (weighted links) to be displayed. (C1)–(C1′) Attention visualization with graph-based drawing for bidirectional relations across sentence words, as encoded in the specified attention layer and head. Users can select a word (e.g., transdermal) and highlight its source sentence and attention pattern, e.g., a drug name (buspirone) and a related disease name (adhd) are gaining strong attentions (C1); or let the word color reflect sentence membership (C1′). (C2) Alternative attention visualization with heatmap. (D) Query panel with a ranked list of attention patterns, which are similar to the attention pattern of the user selected sentence in (C1). The source sentence/document of a queried attention pattern is presented for IR purposes. Additionally, (E1) Word clouds to summarize the most salient information captured by each of the 8 attention heads (from the top attention layer) across the IR corpus. Words are colored by part-of-speech (POS) tags. (E2) Bipartite graphs to summarize the most salient relations captured by the attention heads across the IR corpus.

corpus [T4]. More specifically, we include a corpus navigation panel (Fig. 3A) to select a document or a set of enclosed sentence(s) for exploration. On the sentence/document attention exploration panel (Fig. 3B–C), an attention pattern can be selected to query other documents containing a similar attention pattern. The queried results, including a ranked list of similar attention patterns and the source sentences/documents, are arranged on the query panel (Fig. 3D). These together serve for the domain question to *exploit USE and its essential attention mechanism to promote real-world IR.* For the rest of this section, we present detailed descriptions and rationales of USEVis's components.

5.1. Attention visualization

USE's self-attention mechanism attends (connects, associates) every sentence word to all others, with attention weights indicating the connection strength. Thus for each attention head, the attention pattern, i.e., connections across all sentence words, can be formulated as a directed weighted graph. In this graph, words are represented as nodes, and attentions are represented as weighted edges, with the edge direction indicating the flow of attentions. As our primary visual component to visualize and interpret attention patterns, a visualization of such a node-link graph can illustrate the overall attention pattern with all sentence words involved, which is different from some existing visual designs that mainly focused on attentions pertained to individual word(s) (Vaswani et al., 2017; Vig, 2019b,a). Moreover, the graph topology (e.g., cluster or community) would allow us to depict semantic and syntactic properties constituted among multiple words, not being limited to pairwise relations. Also, the most salient information across an attention pattern could also be highlighted by graph centrality properties. On the other hand, a Sankey visualization which is advantageous of revealing information flows, would not serve for the afore-mentioned purposes. A matrix design with a sorting mechanism could be an alternative to depict an overall attention pattern and reveal sophisticated relations among words, but the intuitiveness and scalability can be limited. Please note we consider matrix visualization preserving the word order as our secondary visual component, and more details are addressed at a later part of this section.

More specifically, graph drawing places nodes and edges in a 2D space and displays relational information with the effective and intuitive spatial channel. With a force-directed graph drawing algorithm, such as ForceAtlas2 (Jacomy et al., 2014), strongly connected nodes are placed closer together, and some cluster/community patterns are exposed as visual densities. In this sense, we can reveal strongly connected words by distances or established clusters, e.g., as shown in Fig. 4 left, four words *attention, deficit, hyperactivity, disorder* constituting a semantic concept of disease name, are closely placed and form a network community. Besides, with graph topology, node centrality also indicates node importance, e.g., as shown in Fig. 4 middle, a salient keyword receiving many attentions is displayed with a high centrality and intuitively reflected in the graph drawing. Additionally, salient functional words can also illustrate the sentence structure, as shown in Fig. 4 right. We also resize nodes by their in-degree (received attentions) and render edge thickness by the attention weight. As a result, the graph drawing not only presents an overview with all sentence words involved, but also eases an exploration of interesting linguistic properties and salient information. Furthermore, the advantageous spatial placement also allows us to compare attention patterns across different sentences, without being restricted by the wording or word order.

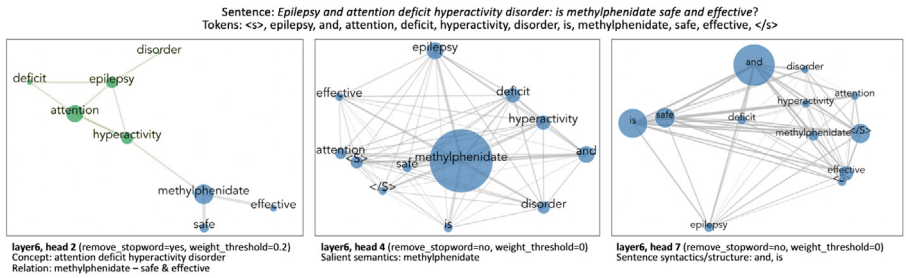

Fig. 4. Graph-based visualizations of sentence-level attention patterns, revealing domain concepts and relations encoded in attention heads.

We also enable interactive adjustments to better support the interpretation and exploration. Specifically, we arrange a *stopword list* which covers common stopwords and special tokens (e.g., ⟨s⟩ and ⟨/s⟩). Users can specify whether to remove tokens from the stopword list or not. In our design session, domain experts commented that removing stopwords can help them better focus on some key concepts, while preserving stopwords can be useful to observe sentence structures or some trivial but interesting information. Besides, users can specify a weight threshold to filter out attentions bearing small weights thus focus on the strongest attentions inferred by USE. The weight threshold also eliminates visual clutters (e.g., hairballs) and highlights the most important network structure, such as communities in a sparsified graph (Fig. 4 left). In addition, when mousing over a node, users can also specify to highlight incoming edges indicating received attentions and/or outgoing edges indicating paid attentions. This makes our visualization provide a comparable view of showing attentions pertained to individual word(s). Finally, users can switch among different attention layers and attention heads, thus examine different attention patterns and explore different aspects of a given sentence.

Alternatively, we also visualize the directed weighted graph of an attention pattern as a heatmap, which is depicted as a colored matrix arranging sentence words as rows and columns while preserving the original word order. For example, a sentence with n words corresponds to an $n \times n$ matrix, where cell (i, j) indicates the attention from word i to word j, with the attention weight encoded by cell color and intensity. In other words, row i reflects how word i pays (distributes) attentions to others, and column j reflects how word j receives attentions from others. Thus the matrix is asymmetric as the attentions a word pays and receives can be different. The advantage of heatmap visualization is to preserve the word order, making it easier to examine a particular sentence and compare its multiple attention patterns from multiple layers and heads. Fig. 5 illustrates the heatmap visualization for a sentence, given its attention patterns across 2 layers and 8 heads.

5.2. Summarizing attentions for multi-head

Given an attention head, besides toggling different sentences and exploring the attention patterns (supported by Section 5.1), we also present automatic summaries with the most salient information captured by the attention head across different sentences. To mitigate distractions from unbounded diversity, we can have a controlled analysis with sentences under a similar topic. In this sense, per a set of specified sentences, we provide visual summaries for each attention head, and enable side-by-side comparisons across different attention heads.

We present two visual summaries for each attention head: a word cloud presents the most salient keywords (the most attended words), and a bipartite graph presents the most salient relations (the most weighted attentions). For the word cloud, we take the k (e.g., 20) most frequent salient words, i.e., words that are frequently emphasized by a given attention head across different sentences. We have word size reflects a word's sentence frequency and word color indicates a word's part-of-speech (POS) tagging, which is commonly used for word-category disambiguation based on its linguistic definition and context. As shown in Fig. 3 E1, the word clouds provide rapid summaries of the 8 attention heads on layer6, with respect to a group of clinical sentences. For example, while head1-6 all distribute attentions to conceptually meaningful words, head1 is more likely to capture patient population (e.g., *children, boys, adolescents, adults*). On the other hand, head7-8 tend to encode syntactic information by distributing attentions to functional words (e.g., *with, of, in, on, and*). Besides, the dominant color(s) of a word cloud also help probe an immediate linguistic role of an attention head, as well as enable a rapid side-by-side comparison. For example, the word clouds of head1-6 are dominated by a color indicating *nouns*, whereas head7-8 are dominated by a color indicating *prepositions or subordinating conjunctions*.

For the bipartite graph, we apply a similar scheme to select the top k salient relations (between word pairs) bearing the highest sentence frequency. The bipartite graph puts source words (paying attention) and target words (receiving attention) into two disjoint sets, and merges repeated words within each set. It provides an additional view for some of the most important relational information captured by an attention head. As shown in Fig. 3 E2, for example, we can further observe that head1 not only derives attentions from disease (*disorder*) to patient (*children*), but also from treatment (*medication, methylphenidate, stimulant, treatment*) to patient (*children*). This might seem similar to the existing work of using a bipartite graph to illustrate self-attention within a sentence, but our bipartite graph is different in (1) summarizing salient relations across multiple sentences, and (2) probing the role of an attention head.

5.3. Synthesizing attentions for documents

While the components of Sections 5.1 and 5.2 would give users insights for model interpretation and confidence in model utilization, starting from this section, we introduce additional components to facilitate exploiting USE and the attentions in real-word IR towards documents.

Based on sentence-level (coherent context) attentions from USE, we synthesize attentions for a document (diversified contexts) consisting of multiple sentences. As described in Section 5.1, a sentence-level attention pattern is formulated as a directed weighted graph. For a document $d = \{s_1, \ldots, s_n\}$, where $s_1, \ldots, s_n$ are enclosed sentences, we consider an union of all tokens such that $tokens(d) = tokens(s_1) \cup \cdots \cup tokens(s_n)$, and a combined set of all attentions such that $attentions(d) = attentions(s_1) \cup \cdots \cup attentions(s_n)$. In this sense, an overlapping token (i.e., a token appearing in multiple sentences) performs as a joint node connecting multiple sub-graphs corresponding to

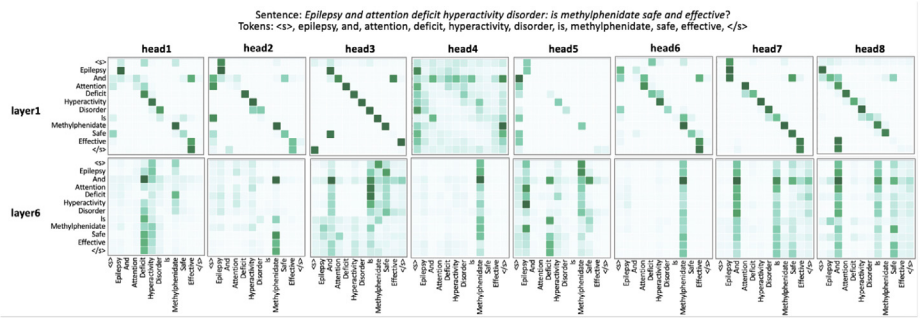

Fig. 5. Heatmap-based visualizations of sentence-level attention patterns across multi-layer (for evolution) and multi-head (for diversity).

multiple sentences; and for overlapping attentions (i.e., a pairwise relation appearing in multiple sentences), we merge them by adding up the weights. This synthesized graph is then visualized by the same scheme of Section 5.1, as shown in Fig. 6. In addition, considering the scalability and heterogeneity, we supplement the following features to better visualize and explore attentions lie in a document: (1) Given a document with multiple sentences, users can select any subset of sentences to synthesize attentions. According to domain experts' feedback, this improves both flexibility and visual effectiveness when particular sentences are preferred per prior knowledge. (2) For graph-based visualization, we color nodes (words) based on their source sentences. For a word appearing in multiple sentences, we visualize it as a pie with multiple colors. (3) To accommodate a larger number of nodes and make their labels visible, we provide sliding bars to increase or decrease node and label size. When scaling, the node size remains proportional to the weighted node degree.

As shown in Fig. 6, a synthesized graph tends to establish clustering patterns aligning with sentences. Besides, some hub nodes can be in correspondence with salient information associated with different contexts. For example, a treatment method (e.g., *methylphenidate* in Fig. 6A) can attend to patient information in one sentence, while being attended by outcome metrics in another sentence. Therefore, such a visualized graph not only depicts a document's natural structure, but also highlights the salient or multi-faceted information analogous to document topics, but in a more descriptive way.

5.4. Attention representation

We generate comparable representations of attention patterns to support interactive query of similar ones, which further prepares us for exploiting attentions in IR.

As an attention pattern is a directed weighted graph (network), an effective representation should encode the network structure, such as the connectivity among nodes or attentions among words. Motivated by the existing work of random (network) walk and node2vec (Grover and Leskovec, 2016), we propose net2vec to learn the representations of a network for an attention pattern. Under this notion, we conduct random walks to sample paths from a network, such that each path consists of a sequence of nodes. We then treat each path as a context of words which are associated by attentions; and all sampled paths together can be an approximation of the entire network, thus the overall attention pattern. To learn a net2vec representation based on a batch of paths/contexts, we feed the sampled paths to the Paragraph Vector model (i.e., doc2vec, Le and Mikolov, 2014). This is analogous to using the word2vec model to learn a node2vec representation (Grover and Leskovec, 2016), utilizing distributional context information.

There exist several hyper-parameters with net2vec. For random walk, we consider a directed weighted network so that the probability in node sampling is proportional to the edge weights. We sample start nodes based on their activities (out-degree), and set the maximum path length as 10. We also reduce revisiting the same nodes by setting $p = 2$ and $q = 0.5$, where p and q are parameters from the random walk algorithm (Grover and Leskovec, 2016) to balance the depth-first search and breadth-first search. For doc2vec (Le and Mikolov, 2014), we examine different window sizes and select a value of 1, meaning only 1-hop distance or directly connected words are considered to be related. This can be explained by the nature of attentions which attend a word to another — transferability might not be applicable. Finally, we take an embedding size (dimensionality of net2vec) of 300, while doc2vec is robust to this parameter.

With net2vec representations of attention patterns, we are able to make comparisons among attention patterns, which can be cross sentences and heads. For a user specified attention pattern of interest (e.g., based on Sections 5.1–5.3), we can provide a list of similar attention patterns by computing the cosine similarity.

5.5. Attention in information retrieval

In Sections 5.1–5.4, we have built components preparing us for exploiting interpretable and comparable attentions. Here we present a typical workflow that USEVis can support: (1) Navigate through a corpus of documents, select a seed document or any subset of enclosed sentences. (2) Explore visualizations of the synthesized attention patterns across attention layers and heads, and identify an (sentence-level) attention pattern of interest. (3) Retrieve similar attention patterns along with their source sentences and documents across the corpus, thus identify other documents enclosing desired semantics.

Accordingly, as shown in Fig. 3, USEVis mainly consists of a navigation panel, an attention panel, and a query panel. The **navigation panel** displays a 2-level hierarchy: document (item) and sentence. As suggested by domain experts, we provide an auxiliary feature to locate sentences via lexical keyword search. The **attention panel** follows the attention visualization described in Sections 5.1 and 5.3. Once sentences are selected, users can specify the attention layer and head, switch between graph-based and heatmap-based visualizations, and customize nodes (words) and edges (attentions) to be displayed. By default, all words and attentions are preserved. With a graph-based visualization, users can hover a node to highlight its incoming edges, outgoing edges, or both; users can also click a node and all associated nodes will be highlighted. Once a user selects an attention pattern, the **query panel** displays the k most similar attention patterns in a ranked order. For each recommended attention pattern, we display the source sentence and document in correspondence. Additionally, we have an **attention head summary panel**, as described in Section 5.2, to provide two conveniences: (1) advice the selection of an attention head which is more likely to encode properties of interest across the corpus, and (2) provide a quick topic overview of the corpus at no extra cost.

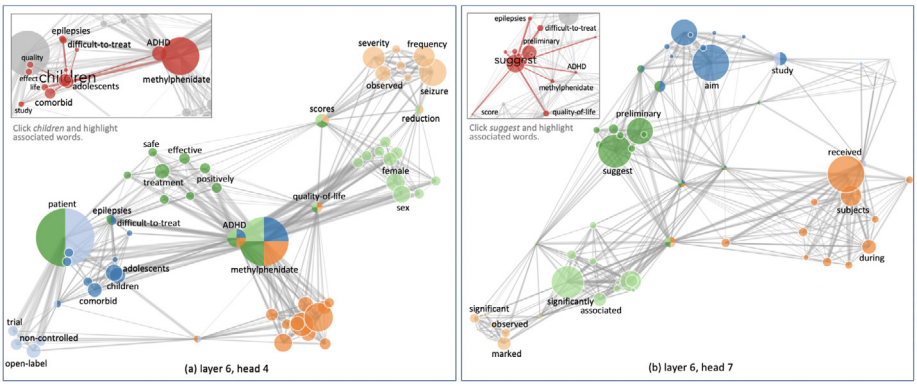

Fig. 6. Synthesized graph visualization of attention patterns for a document consisting of multiple sentences, assisting in document understanding.

Generating representations for attention patterns with net2vec

Step1: Take an attention pattern as a directed weighted network;
Step2: Perform random walk and sample paths from the network, with $p = 2, q = 0.5$;
Step3: Feed sampled paths (aka., sequences of words) to doc2vec, with $window\_size = 1, dimensionality = 300$;
Step4: Obtain the paragraph or document embedding produced by doc2vec as the network or attention representation.

6. Case study and expert feedback

Two domain experts (in text mining and clinical IR) participated in our case studies. We took a bottom-up procedure in order to (1) Gain insights and confidence in the fundamental attention mechanism by interpreting how meaningful semantics or syntactics are encoded by attention heads, with respect to individual sentences in a manageable and intuitive manner [T1]. (2) Gain further insights into the multi-head setting of the attention mechanism, via summarized information and side-by-side comparisons across different attention heads, with respect to a group of sentences under a similar topic [T2]. After this bottom-up procedure, we move forward to (3) Exploit interpretable attention patterns at the document-level [T3] and facilitate identifications of relevant documents enclosing attention patterns of domain or task interest, with respect to an IR corpus [T4]. As suggested by domain experts, throughout the case studies, we use a document corpus consisting of 851 clinical trials which are related to the drug effectiveness of attention deficit hyperactivity disorder (ADHD). Domain experts are interested in probing how domain concepts (such as patient, disease, treatment) and relations (such as patient-and-disease, disease-and-treatment, treatment-and-outcome) are encoded by attention heads, and utilizing recognized attention patterns encoding interesting concepts/relations to retrieve documents that are relevant to the IR information need.

6.1. Domain concepts and relations

With the ADHD corpus, domain experts identified a sample sentence from a document title: *Epilepsy and attention deficit hyperactivity disorder: is methylphenidate safe and effective?* (Gross-Tsur et al., 1997). This sentence is selected as it contains a multi-world concept of disease name (*attention deficit hyperactivity disorder*), an important drug name (*methylphenidate*), a relation between a drug and its outcome (*safe and effective*), and an interesting syntactic sentence structure (*is … ?*).

By examining graph-based visualizations of attention patterns from head1–head8 of the top layer (layer 6), experts recognized that different attention heads tend to depict this sentence from different perspectives. Specifically, as shown in Fig. 4, experts highlighted that: (1) head2 encodes mutual associations and forms a community across the multiple words of *attention*, *deficit*, *hyperactivity*, and *disorder*, successfully capturing the domain concept of a disease name; besides, head2 also encodes the relation from *methylphenidate* to *safe* and *effective*, establishing the domain relation of treatment-and-outcome. (2) head4 and head6 encode similar patterns of distributing the majority of attentions from (almost) all words to one word *methylphenidate*, pinpointing the important drug name which performs as the most salient information of the sentence. (3) head7 and head8 derive more attentions to functional words such as *and* and *is*, which indicate the sentence type as questioning. One of the experts commented that such syntactic insights are also beneficial to IR. For example, when a research question is asked, a sentence with the *is … ?* structure will be valued. Besides, there are also trivial patterns or noisy information encoded in some attentions, for example, head1 attends *is*, *safe*, and *effective* to *deficit*, which is an improper relation between outcome and disease. After screening 20 sample sentences in total, the domain expert was impressed by how most attention heads (especially from the top layer) can encode domain concepts, relations, and even sentence structures in an insightful way. They gained better confidence that the attention mechanism is promising to capture essential sentence meanings, thus enabling the generation of high-quality sentence embeddings. Worth mentioning, the experts also made hypotheses on the roles of head2, head4, and head7, based on the afore-descriptions.

In addition, as heatmap is advantageous when comparing different attention patterns lie in the same sentence, domain experts also looked into a collection of heatmap across attention layers and heads. In Fig. 5, we present the bottom layer (layer1) and the top layer (layer6). We found that for the bottom layer, sentence words are likely to attend to (or be attended by) themselves or immediate neighbors. After evolution through

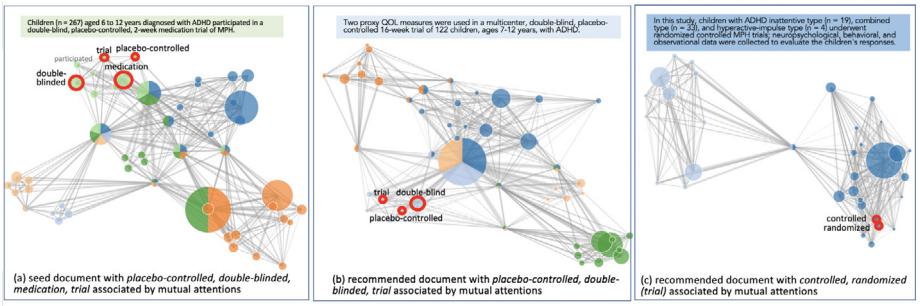

Fig. 7. Identification of a relevant attention pattern from a seed document (a), and retrieval of similar attention patterns across corpus for IR (b–c).

middle layers (layer2 to layer5), attentions become more likely to associate words with a larger distance and capture more diverse patterns. The top layer, as built upon context information accumulated from lower layers, is prone to encode higher-level linguistic patterns. Besides, by making side-by-side comparisons across attention heads, one domain expert commented it to be helpful for her to gain a quick sense about the diverse sentence aspects captured by different heads.

6.2. Attention heads and linguistic properties

This use case is designed to understand potential roles pertained to attention heads. Before going to visual summaries of attention heads, domain experts decided to start with a more controlled analysis by utilizing sentence pairs carrying similar wording, semantics, or syntactics.

Pair 1: *Is methylphenidate safe and effective for children with attention deficit?* vs. *Is methylphenidate effective and safe for children with attention deficit?* With graph-based and heatmap-based visualizations, domain experts agreed that the attention heads tend to produce consistent patterns across the two sentences, although differences are also noted. In particular, without being interfered by the word order, head3 constantly establishes strong relations between *safe* and *attention*, and between *effective* and *attention deficit*.

Pair 2: *Is methylphenidate safe and effective for children with attention deficit?* vs. *Is methylphenidate safe and effective for adults with attention deficit?* With head1, *children* and *adult* receive similar attentions from others, including the strongest attention from *methylphenidate*. With head5, both *children* and *adult* pay exclusively strong attention to *methylphenidate*. However, head3 is an exception: *children* receives attentions from *is methylphenidate safe for*, but *adult* does not.

Pair 3: *Epilepsy and attention deficit hyperactivity disorder: is methylphenidate safe and effective?* vs. *Headache and chronic recurring migraine: are pain relievers effective and safe?* We had sentence pairs with similar syntactics but different wordings. While most attention heads encode different patterns for these two sentences, head7 and head8 consistently capture the sentence functional word *is* and *are*.

At this point, we are able to identify some attention heads tending to encode consistent linguistic properties. However, with increased sentence complexity, some attention heads are deployed to capture more diverse or variant information. Domain experts then moved to text clouds and bipartite graphs that are summarized from more sentences under a similar topic, i.e., 851 document titles from the ADHD corpus, for each of the attention heads. Although variations can exist, they are interested in observing the overall most salient information encoded by each head. One expert appreciated how the word cloud colors help gain a rapid sense that head7 and head8 focus on functional words and syntactic patterns; furthermore, head4 tends to have a full focus on nouns and discretely salient words; and head2 highlights both nouns and functional words, tending to capture mutual relationships across multiple words. Furthermore, by looking into the bipartite graphs, head4 tends to derive attentions to *methylphenidate* and head2 tends to form mutual relations among *attention*, *deficit*, *hyperactivity*, and/or *disorder*. Domain experts considered these provide positive support to their hypotheses made in Section 6.1. They also commented on the meaningfulness of these investigations to help understand the multi-head mechanism and some dedicated roles in potential.

6.3. Attention in information retrieval

With obtained insights, one expert utilized USEVis to conduct an IR task with a purpose to evaluate the potential of exploiting interpretable and comparable attention patterns to facilitate the identification of relevant documents. As a full IR task can be time-consuming, we selected 20 documents (clinical trials) with a mixture of relevant and irrelevant ones from the ADHD corpus. We use each clinical trial's abstract, which is the most informative part consisting of around 10 sentences, as the document instance.

As the selected documents are mostly about the methylphenidate treatment for ADHD, the expert further elaborated her interest to identify relevant documents based on the study design (*controlled clinical trial*), patient (*children*), and treatment outcome (*quality of life, metric, score, etc.*). She started with a random document, selected a subset of 6 sentences via the navigation panel, and triggered a synthesized graph visualization (Fig. 6a). She commented on the effectiveness of interpreting salient or relational information. Specifically, she rapidly learned that head4 (of the top layer) is the most interesting one. As shown in Fig. 6a, with head4, (1) *ADHD*, *methylphenidate*, and *quality of life* are placed with a high centrality; (2) the established clusters also highlight key information of patient population groups (*children with epilepsy*), study type (*non-controlled trial*), outcome (*safe and effective*), etc. respectively; (3) the graph structure across clusters and centralized nodes also reveal domain relations, for example, *ADHD* and *methylphenidate* are closely attended to each other reflecting the disease-and-treatment relation, and *ADHD* and *methylphenidate* are also associated with the clusters indicating patient and outcome. The expert was satisfied with how head4 encodes important concepts and relations, and our visual design is effective for her to reach an intuitive understanding [T1, T3]. In contrast, as shown in Fig. 6b, head7 highlights many functional words and tends to encode syntactic information. After re-visiting the visual summaries of different attention heads, the expert was impressed by the diversified properties being encoded. She agreed with the necessity to explore and gain more insights into the multi-head mechanism, thus advising a more favorable utilization of attentions once the black box is opened [T2].

Furthermore, the expert was also satisfied that USEVis can efficiently suggest the document type indicating the clinical study design. For example, as shown in Fig. 7(a), *open-label*, *non-controlled*, and *trial* are associated by attentions and form a domain concept. Therefore, the expert was able to quickly skip a few documents until identifying a seed document with a desired type of *controlled trial* [T1, T4]. With this seed document, she then specified a sentence-level attention pattern which encodes the concept of *double-blind placebo-controlled medication trial* (as shown in Fig. 7(a)), and leveraged it to retrieve other documents with similar attention patterns. Powered by comparable attention representations, Fig. 7(b)(c) illustrate the recommended documents, which enclose concepts of *double-blind placebo-controlled trial* and *controlled randomized (trial)* respectively. Following this procedure, the expert was able to identify other relevant documents with a desired type. She commented that the attention patterns allowed her to understand and focus on a particular document facet or topic, thus retrieving interesting documents more precisely and efficiently. She embraced the potential to promote IR applications [T4].

Overall, Domain experts considered USEVis as an interesting and innovative effort that not only helps them gain insights and confidence in applying the attention-driven model to domain applications, but also provides a platform for them to conduct IR tasks leveraging attention patterns. They are satisfied with the design and the overall usefulness of USEVis. They are also impressed by the visual analytics results accumulated during an exploration process. Meanwhile, they suggested that the accuracy of retrieving similar attention patterns could be further improved. As there are multiple influential factors, we plan to address this in a future effort.

7. Discussion

What is Encoded in Attention? Our visual analytics system allowed domain experts to interpret, explore, and gain insights into attentions learned by USE. With sentences and documents from clinical corpus, we found established attention patterns are able to capture interesting domain concepts, such as patient, disease, treatment, outcome, clinical study design, etc., by grouping multiple words (e.g., *attention, deficit, hyperactivity, disorder*) constituting a standard concept via mutual attentions, or highlighting a few salient words (e.g., *methylphenidate*) via (exclusively) strong attentions. Meanwhile, there are also attention patterns connecting relational words and forming meaningful domain relations, such as patient-and-disease, disease-and-treatment, treatment-and-outcome, etc. Besides, we also notice some attention patterns accounting for the sentence structure, via emphasizing functional words such as *is*, *are*, *and*, etc. The attentions encoding such semantic or syntactic properties perform as building blocks of high-quality sentence embedding, thus help explain the success of USE. On the other hand, there also exist attention patterns corresponding to trivial information or inaccurate semantics/syntactics, which might in turn help diagnose an attention-based encoder. Under this notion, regarding the interpretability of attention, our observations suggested that although not always, attention can be interpretable with respect to the attention patterns revealing concepts and relations of IR interest.

Multi-layer and Multi-head Attention. The visualization of attention patterns across layers and heads would also explain the design rationale of a multi-layer and multi-head attention-based neural model, such as USE and BERT. Specifically, we found attention patterns through layers, i.e., from lower layers to upper layers, demonstrated an evolution from capturing simple syntactics (e.g., a word attending to itself or its immediate neighbors) to encoding more complicated syntactic (e.g., long-term dependency and sentence structure) and meaningful semantics (e.g., concepts and relations). For each layer, the multiple attention heads further depict text meanings from different perspectives. Our domain experts confirmed the usefulness of multi-head attentions in approaching a multi-faceted sentence or document. This is specifically promising in IR applications when an analytical focus, particular domain interest, or task specification is expected. In addition, being inspired by domain experts, the multi-head attention patterns (especially of the top layer) can be analogues to a series of text topics. While topic modeling, such as LDA and NMF, is based on lexical and statistical computation, the attention patterns would be advantageous in encoding complicated syntactics and underlying semantics, as well as leveraging knowledge pre-trained by extensive resources.

From Sentence-level Attention to Document-level Attention. USE and other attention-based neural models generally treat a sentence as a coherent context. In USE, attentions are established across words within a sentence. Our first contribution is to provide interactive visual analytics of sentence-level attentions. Furthermore, with respect to real-world IR towards a corpus of documents, we further contribute to synthesize document-level attentions and extend our visual analytics accordingly. In our iterative design and development process, there were several challenges when synthesizing attentions for a document, including accommodating a large number of tokens into a graph, diversified information from different sentences with unrelated contexts, etc. While we intended to provide a unified system for both sentence-level and document-level exploration, we consider a follow-up study could focus on document-level attentions in particular. In addition, for the recommendation of similar attention patterns, future effort would support different granularities when selecting a seed attention pattern, such as document-level (entire synthesized graph), sentence-level (sub-graph corresponding to a sentence), node-level (node with its local neighborhood), and customization-based (i.e., any sub-graph).

8. Conclusion

We designed and developed a visual analytics system to interpret, explore, and exploit attention patterns learned by an attention-driven neural embedding model, Universal Sentence Encoder. With an IR application-driven view, we mainly focused on attentions patterns encoding linguistic properties (semantics and syntactics) corresponding to domain concepts and relations. Moreover, we examined attention patterns from multiple layers and multiple heads, and shed light on the existence of attention heads encoding consistent properties across different sentences. With respect to domain IR applications in practice, we extended our approach from analyzing sentence-level attentions to document-level attentions. Given a corpus for information retrieval, we supported the retrieval of similar attention patterns, thus facilitate the identification of relevant sentences or documents in correspondence. Putting things together, our visual analytics system enabled intuitive visualization as well as interactive exploration for the aforementioned purposes. We received inspiring findings from use cases and positive feedback from domain experts. We verified the usefulness and effectiveness of our system, and the potential of leveraging state-of-the-art attention models to better support real-world applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Battista, G.D., Eades, P., Tamassia, R., Tollis, I.G., 1994. Algorithms for drawing graphs: An annotated bibliography. Comput. Geom. 4 (5), 235–282.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Chen, J., Zhuge, H., 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4046–4056.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., Hu, G., 2016. Attention-over-attention neural networks for reading comprehension. arXiv preprint arXiv:1607.04423.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Görg, C., Liu, Z., Stasko, J., 2014. Reflections on the evolution of the Jigsaw visual analytics system. Inf. Vis. 13 (4), 336–345.

Gross-Tsur, V., Manor, O., Van der Meere, J., Joseph, A., Shalev, R., 1997. Epilepsy and attention deficit hyperactivity disorder: Is methylphenidate safe and effective? J. Pediatr. 130 (1), 40–44.

Grover, A., Leskovec, J., 2016. Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 855–864.

Henry, N., Fekete, J., 2006. Matrixexplorer: A dual-representation system to explore social networks. IEEE Trans. Vis. Comput. Graphics 12 (5), 677–684.

Henry, N., Fekete, J.-D., McGuffin, M.J., 2007. NodeTrix: A hybrid visualization of social networks. IEEE Trans. Vis. Comput. Graphics 13 (6), 1302–1309.

Henry Riche, N., Fekete, J.-D., 2007. MatLink: Enhanced matrix visualization for analyzing social networks. In: Lecture Notes in Computer Science (Proceedings of the 13th IFIP TC13 International Conference on Human-Computer Interaction, INTERACT'07). pp. 288–302.

Herman, I., Melancon, G., Marshall, M.S., 2000. Graph visualization and navigation in information visualization: A survey. IEEE Trans. Vis. Comput. Graphics 6 (1), 24–43.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Jacomy, M., Venturini, T., Heymann, S., Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 9 (6), e98679.

Ji, X., Shen, H.-W., Ritter, A., Machiraju, R., Yen, P.-Y., 2019. Visual exploration of neural document embedding in information retrieval: Semantics and feature selection. IEEE Trans. Vis. Comput. Graphics 25 (6), 2181–2192.

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors. In: Proceedings of Advances in Neural Information Processing Systems, vol. 28, pp. 3294–3302.

Koren, Y., 2005. Drawing graphs by eigenvectors: Theory and practice. Comput. Math. Appl. 49 (11–12), 1867–1888.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196.

Lin, Z., Feng, M., dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y., 2019. A structured self-attentive sentence embedding. arXiv:1703.03130.

Liu, S., Bremer, P., Thiagarajan, J.J., Srikumar, V., Wang, B., Livnat, Y., Pascucci, V., 2018. Visual exploration of semantic relationships in neural word embeddings. IEEE Trans. Vis. Comput. Graphics 24 (1), 553–562.

Lopez, M.M., Kalita, J., 2017. Deep learning applied to NLP. arXiv preprint arXiv:1703.03091.

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., Qu, H., 2017. Understanding hidden memories of recurrent neural networks. In: Proceedings of 2017 IEEE Conference on Visual Analytics Science and Technology. pp. 13–24.

Mitra, B., Craswell, N., 2017. Neural text embeddings for information retrieval. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, pp. 813–814.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R., 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (4), 694–707.

Park, C., Na, I., Jo, Y., Shin, S., Yoo, J., Kwon, B.C., Zhao, J., Noh, H., Lee, Y., Choo, J., 2019. SANVis: Visual analytics for understanding self-attention networks. arXiv:1909.09595.

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F.B., Wattenberg, M., 2016. Embedding projector: Interactive visualization and interpretation of embeddings. arXiv preprint arXiv:1611.05469.

Strobelt, H., Gehrmann, S., Pfister, H., Rush, A.M., 2018. LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Trans. Vis. Comput. Graphics 24 (1), 667–676.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Vig, J., 2019a. A multiscale visualization of attention in the transformer model. arXiv:1906.05714.

Vig, J., 2019b. Visualizing attention in transformer-based language representation models. arXiv:1904.02679.

Zhao, J., Sun, M., Chen, F., Chiu, P., 2017. BiDots: Visual exploration of weighted biclusters. IEEE Trans. Vis. Comput. Graphics 24 (1), 195–204.