Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach

Saurav Ghosh
Virginia Tech
sauravcsvt@vt.edu

Prithwish Chakraborty
Virginia Tech
prithwi@vt.edu

Emily Cohn
Boston Children's Hospital
emily.cohn@childrens.harvard.edu

John S. Brownstein
Harvard Medical School
john.brownstein@childrens.harvard.edu

Naren Ramakrishnan
Virginia Tech
naren@cs.vt.edu

ABSTRACT

Traditional disease surveillance can be augmented with a wide variety of real-time sources such as, news and social media. However, these sources are in general unstructured and, construction of surveillance tools such as taxonomical correlations and trace mapping involves considerable human supervision. In this paper, we motivate a disease vocabulary driven word2vec model (*Dis2Vec*) to model diseases and constituent attributes as word embeddings from the HealthMap news corpus. We use these word embeddings to automatically create disease taxonomies and evaluate our model against corresponding human annotated taxonomies. We compare our model accuracies against several state-of-the art word2vec methods. Our results demonstrate that *Dis2Vec* outperforms traditional distributed vector representations in its ability to faithfully capture taxonomical attributes across different class of diseases such as endemic, emerging and rare.

CCS Concepts

•**Information systems** → **Data mining;** *Information retrieval;*

Keywords

Disease characterization; Domain-specific word embeddings; word2vec; *Dis2Vec*; HealthMap

1. INTRODUCTION

Traditional disease surveillance has often relied on a multitude of reporting networks such as outpatient networks, on-field healthcare workers, and lab-based networks. Some of the most effective tools while analyzing or mapping diseases, especially for new diseases or disease spreading to new regions, are reliant on building disease taxonomies which can aid in early detection of outbreaks.

CIKM'16 , October 24-November 28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983362

In recent years, the ready availability of social and news media has led to services such as HealthMap [7] which have been used to track several disease outbreaks from news media ranging from the flu to Ebola. However, most of this data is unstructured and often noisy. Annotating such corpora thus requires considerable human oversight. While significant information about both endemic [4, 22] and rare [19] diseases can be extracted from such news corpora, traditional text analytics methods such as lemmatization and tokenization are often shallow and do not retain sufficient contextual information. More involved methods such as topic models are too computationally expensive for real-time worldwide surveillance and do not provide simple semantic contexts that could be used to comprehend the data.

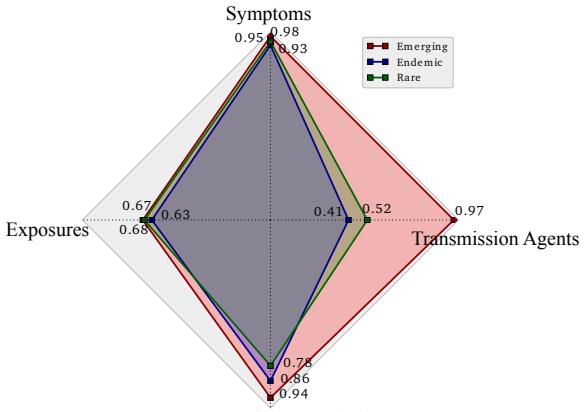


Figure 1: Comparative performance evaluation of disease specific word2vec model (*Dis2Vec*) across the disease characterization tasks for 3 different class of diseases - endemic (blue), emerging (red) and rare (green). The axes along the four vertices represent the modeling accuracy for the disease characterization of interest viz. symptoms, transmission agents, transmission methods, and exposures. The area under the curve for each disease class represent the corresponding overall accuracy over all the characterizations. Best characterization performance can be seen for emerging diseases.

In recent years, several deep learning based methods, such as word2vec and doc2vec, have been found to be promising in analyzing such text corpora. These methods once trained over a representative corpus can be readily used to analyze new text and find semantic constructs (e.g. *rabies*:*zoonotic* = *salmonella*:*foodborne*) which can be useful for automated taxonomy creation. Classical word2vec methods are generally unsupervised requiring no domain information and as such has broad applicability. However, for highly specified

domains (such as disease surveillance) with moderate sized corpus, classical methods fail to find meaningful semantic relationships. For example, while determining the transmission method of salmonella given that rabies is zoonotic (i.e. querying *rabies*:*zoonotic* = *salmonella*:??), traditional word2vec methods such as skip-gram model trained on the HealthMap corpus fail to find a meaningful answer (*saint-paul*).

Motivated by this problem, in this paper we postulate a vocabulary driven word2vec algorithm that can find meaningful disease constructs which can be used towards such disease knowledge extractions. For example, for the aforementioned task, vocabulary-driven word2vec algorithm generates the word *foodborne*, which is more meaningful in the context of disease knowledge extraction. Our main contributions are:

• We formulate *Dis2Vec*, a vocabulary driven word2vec method which is used to generate disease specific word embeddings from unstructured health-related news corpus. *Dis2Vec* allows domain knowledge in the form of pre-specified disease-related vocabulary $\mathcal{V}$ to supervise the discovery process of word embeddings.
• We use these disease specific word embeddings to generate automated disease taxonomies that are then evaluated against human curated ones for accuracies.
• Finally, we evaluate the applicability of such word embeddings over different class of diseases - emerging, endemic and rare for different taxonomical characterizations.

Preview of our results: In Figure 1, we provide a comparative performance evaluation of *Dis2Vec* across the disease characterization tasks for endemic, emerging and rare diseases. It can be seen that *Dis2Vec* is best able to characterize emerging diseases. Specifically, it is able to capture symptoms, transmission methods and transmission agents, with near-perfect accuracies for emerging diseases. Such diseases (e.g. Ebola, H7N9) draw considerable media interest due to their unknown characteristics. News articles reporting emerging outbreaks tend to focus on all characteristics of such diseases - symptoms, exposures, transmission methods and transmission agents. However, for endemic and rare diseases, transmission agents and exposures are better understood, and news reports tend to focus mainly on symptoms and transmission methods. *Dis2Vec* can still be applied for these class of diseases but with decreased accuracy for these under-represented characteristics.

2. RELATED WORK

The related works of interest for our problem are primarily from the field of neural-network based word embeddings and their applications in a variety of NLP tasks. In recent years, we have witnessed a tremendous surge of research concerned with representing words from unstructured corpus to dense low-dimensional vectors drawing inspirations from neural-network language modeling [3, 5, 15]. These representations, referred to as *word embeddings*, have been shown to perform with considerable accuracy and ease across a variety of linguistic tasks [1, 6, 20].

Mikolov et al. [12, 13] proposed skip-gram model, currently a state-of-the-art word embedding method, which can be trained using either hierarchical softmax (SGHS) [13] or the negative sampling technique (SGNS) [13]. Skip-gram

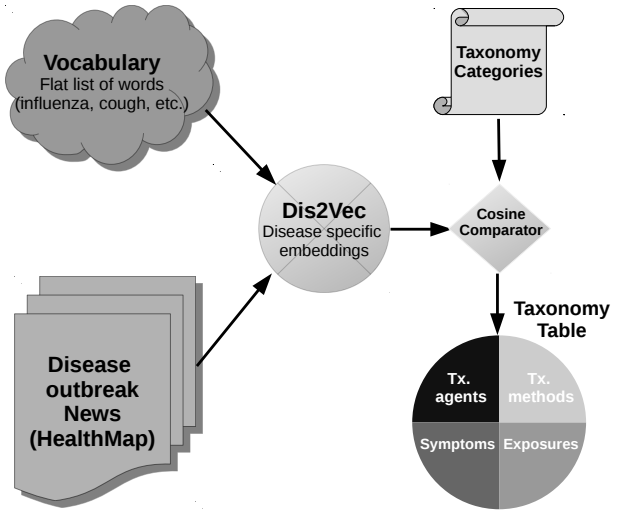


Figure 2: Automated taxonomy generation from unstructured news corpus (HealthMap) and a pre-specified vocabulary ($\mathcal{V}$). *Dis2Vec* inputs these information to generate disease specific word embeddings that are then passed through a cosine comparator to generate the taxonomy for the disease of interest.

models have been found to be highly efficient in finding word embedding templates from huge amounts of unstructured text data and uncover various semantic and syntactic relationships. Mikolov et al. [13] also showed that such word embeddings have the capability to capture linguistic regularities and patterns. These patterns can be represented as linear translations in the vector space. For example, vec(*Madrid*) - vec(*Spain*) + vec(*France*) is closer to vec(*Paris*) than any other word in their corpus [14, 9].

Levy et al. [10] analyzed the theoretical founding of skip-gram model and showed that the training method of SGNS can be converted into a weighted matrix factorization and its objective induces an implicit factorization of a shifted PMI matrix - the well-known word-context PMI matrix [2, 21] shifted by a constant offset. In [11], Levy et al. performed an exhaustive evaluation showing the impact of each parameter (window size, context distribution smoothing, subsampling of frequent words and others) on the performance of SGNS and other recent word embedding methods, such as GLoVe [16]. They found that SGNS consistently profits from larger negative samples (> 1) showing significant improvement on various NLP tasks with higher values of negative samples.

Previous works on *neural embeddings* (including the skip-gram model) define the contexts of a word to be its linear context (words preceding and following the target word). Levy et al. [8] generalized the skip-gram model and used syntactic contexts derived from automatically generated dependency parse-trees. These syntactic contexts were found to capture more functional similarities, while the bag-of-words nature of the contexts in the original skip-gram model generates broad topical similarities.

3. MODEL

3.1 Problem Overview

Disease taxonomy generation is the process of tabulating characteristics of diseases w.r.t. several pre-specified cat-

egories such as symptoms and transmission agents. Table 1 gives an example of taxonomy for three diseases viz. an emerging disease (H7N9), an endemic disease (avian influenza) and a rare disease (plague). Traditionally, such taxonomies are human curated - either from prior expert knowledge or by combining a multitude of reporting sources. News reports covering disease outbreaks can often contain disease specific information, albeit in an unstructured way. Our aim is to generate automated taxonomy of diseases similar to Table 1 using such unstructured information from news reports. Such automated methods can greatly simplify the process of generating taxonomies, especially for emerging diseases, and lead to a timely dissemination of such information towards public health services. In general, such disease related news corpus is of moderate size for deep-learning methods and as explained in section 1, unsupervised methods often fail to extract meaningful information. Thus we incorporate domain knowledge in the form of a flat-list of disease related terms such as disease names, possible symptoms and possible transmission methods, hereafter referred to as the vocabulary $\mathcal{V}$. Figure 2 shows the process of automated taxonomy generation where we employ a supervised word2vec method referred to as *Dis2Vec* which takes the following inputs - (a) the pre-specified disease vocabulary $\mathcal{V}$ and (b) unstructured news corpus $\mathcal{D}$ and generates embeddings for each word (including words in the vocabulary $\mathcal{V}$) in the corpus. Once word embeddings are generated, we employ a cosine comparator to create a tabular list of disease taxonomies similar to Table 1. In this cosine comparator, to classify each disease for a taxonomical category, we calculate the cosine similarities between the embedding for the disease name and embeddings for all possible words related to that category. Then, we sort these cosine similarities (in descending order) and extract the words (higher up in the order) closer to the disease name hereafter referred to as top words found for that category. For example, to extract the transmission agents for *plague*, we calculate the cosine similarities between the embedding for the word *plague* and the embeddings for all possible terms related to transmission agents and extract the top words by sorting the terms w.r.t. these similarities. We can compare the top words found for a category with the human annotated words to compute the accuracy of the taxonomy generated from word embeddings. In the next 2 subsections, we will briefly discuss the basic word2vec model (skip-gram model with negative sampling) followed by the detailed description of our vocabulary driven word2vec model *Dis2Vec*.

Table 1: Human curated disease taxonomy for three diseases from three different class of diseases (endemic, emerging, and rare).

Disease	Transmission methods	Transmission agents	Clinical symptoms	Exposures
Avian influenza (endemic)	zoonotic	domestic animal, wild animal	Fever, cough, sore throat, diarrhea, vomiting	animal exposure, farmer, market, slaughter
H7N9 (emerging)	zoonotic	domestic animal	Fever, cough, pneumonia	farmer, market, slaughter, animal exposure
Plague (rare)	vectorborne, zoonotic	flea, wild animal	Sore, fever, headache, muscle ache, vomiting, nausea	animal exposure, veterinarian, farmer

3.2 Basic Word2vec Model

In this section, we present a brief description of *SGNS* - the skip-gram model introduced in [12] trained using the negative sampling procedure in [13]. The objective of the skip-gram model is to infer word embeddings that will be relevant for predicting the surrounding words in a sentence or a document. It is to be noted that the skip-gram model can also be trained using Hierarchical Softmax method as shown in [13].

The inputs to the skip-gram model are a corpus of words $w \in \mathcal{W}$ and their corresponding contexts $c \in \mathcal{C}$ where $\mathcal{W}$ and $\mathcal{C}$ are the word and context vocabularies. In *SGNS*, the contexts of word w_i are defined by the words surrounding it in an L-sized context window $w_{i-L}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+L}$. In order to convert the corpus $\mathcal{D}$ of unstructured news reports into a collection of observed (w, c) pairs, the textual content of each news report is processed to generate a set of unique terms or words and then the contexts of each such term are extracted by identifying the words surrounding it in an L-sized window. The notation $\#(w, c)$ represents the number of times the pair (w, c) occurs in $\mathcal{D}$. Therefore, $\#(w) = \sum_{c \in \mathcal{C}}(w, c)$ and $\#(c) = \sum_{w \in \mathcal{W}}(w, c)$ where $\#(w)$ and $\#(c)$ are the total number of times w and c occurred in $\mathcal{D}$. Each word $w \in \mathcal{W}$ corresponds to a vector $\mathbf{w} \in R^T$ and similarly, each context $c \in \mathcal{C}$ is represented as a vector $\mathbf{c} \in R^T$, where T is the dimensionality of the word or context embedding. The entries in the vectors are the latent parameters to be learned.

SGNS tries to maximize the probability whether a single word-context pair (w, c) was generated from the observed corpus $\mathcal{D}$. Let $P(\mathcal{D} = 1|w, c)$ refers to the probability that (w, c) was generated from the corpus, and $P(\mathcal{D} = 0|w, c) = 1 - P(\mathcal{D} = 1|w, c)$ the probability that (w, c) was not. The objective function for a single (w, c) pair is modeled as:

$$P(\mathcal{D} = 1|w, c) = \sigma(\mathbf{w} \cdot \mathbf{c}) = \frac{1}{1+e^{-\mathbf{w} \cdot \mathbf{c}}} \quad (1)$$

where $\mathbf{w}$ and $\mathbf{c}$ are the T-dimensional latent parameters or vectors to be learned.

The objective of the negative sampling is to maximize $P(\mathcal{D} = 1|w, c)$ for observed (w, c) pairs while maximizing $P(\mathcal{D} = 0|w, c)$ for randomly sampled *negative* contexts (hence the name *negative sampling*), under the assumption that randomly selecting a context for a given word will tend to generate an unobserved (w, c) pair. *SGNS*'s objective for a single (w, c) observation is then:

$$\log \sigma(\mathbf{w} \cdot \mathbf{c}) + k \cdot \mathbb{E}_{c_N \sim P_{\mathcal{D}}}[\log \sigma(-\mathbf{w} \cdot \mathbf{c_N})] \quad (2)$$

where k is the number of *negative* samples and c_N is the sampled context, drawn according to the smoothed unigram distribution $P_{\mathcal{D}}(c) = \frac{\#(c)^\alpha}{\sum_c \#(c)^\alpha}$ where $\alpha = 0.75$ is the smoothing parameter. $\mathbb{E}$ represents the expectation term.

The objective of $SGNS$ is trained in an online fashion using stochastic gradient updates over the observed pairs in the corpus $\mathcal{D}$. The global objective then sums over the observed (w, c) pairs in the corpus:

$$l_{SGNS} = \sum_{(w,c)\in\mathcal{D}} \Big(\log \sigma(\mathbf{w}\cdot\mathbf{c}) + k \cdot \mathbb{E}_{c_N \sim P_{\mathcal{D}}}[\log \sigma(-\mathbf{w}\cdot\mathbf{c_N})] \Big) \tag{3}$$

Optimizing this objective will have a tendency to generate similar embeddings for observed word-context pairs, while scattering unobserved pairs in the vector space. Intuitively, words that appear in similar contexts or tend to appear in the contexts of each other should have similar embeddings.

3.3 Disease Specific Word2vec Model (*Dis2Vec*)

In this section, we introduce *Dis2Vec*, a disease specific word2vec model whose objective is to generate word embeddings which will be useful for automatic disease taxonomy creation given an input unstructured corpus $\mathcal{D}$. We used a pre-specified disease-related vocabulary $\mathcal{V}$ (domain information) to guide the discovery process of word embeddings in *Dis2Vec*. The input corpus $\mathcal{D}$ consists of a collection of (w, c) pairs. Based on $\mathcal{V}$, we can categorize the (w, c) pairs into three types as shown below:

- $\mathcal{D}_{(d)} = \{(w, c) : w \in \mathcal{V} \wedge c \in \mathcal{V}\}$, i.e. both the word w and the context c are in $\mathcal{V}$
- $\mathcal{D}_{(\neg d)} = \{(w, c) : w \notin \mathcal{V} \wedge c \notin \mathcal{V}\}$, i.e. neither the word w nor the context c are in $\mathcal{V}$
- $\mathcal{D}_{(d)(\neg d)} = \{(w, c) : w \in \mathcal{V} \oplus c \in \mathcal{V}\}$, i.e. either the word w is in $\mathcal{V}$ or the context c is in $\mathcal{V}$ but both cannot be in $\mathcal{V}$

Therefore, the input corpus $\mathcal{D}$ can be represented as $\mathcal{D} = \mathcal{D}_{(d)} + \mathcal{D}_{(\neg d)} + \mathcal{D}_{(d)(\neg d)}$. Each of these categories of (w, c) pairs needs special consideration while generating disease specific embeddings.

3.3.1 Vocabulary Driven Negative Sampling

The first category ($\mathcal{D}_{(d)}$) of (w, c) pairs, where both w and c are in $\mathcal{V}$ ($w \in \mathcal{V} \wedge c \in \mathcal{V}$), is of prime importance in generating disease specific word embeddings. Our first step in generating such embeddings is to maximize $\log \sigma(\mathbf{w}\cdot\mathbf{c})$ in order to achieve similar embeddings for these disease word-context pairs. Apart from maximizing the dot products, following classical approaches [13], negative sampling is also required to generate robust embeddings. In *Dis2Vec*, we adopt a vocabulary ($\mathcal{V}$) driven negative sampling for these disease word-context pairs. In this vocabulary driven approach, instead of random sampling we sample negative examples (c_N) from the set of non-disease contexts, i.e. contexts which are not in $\mathcal{V}$ ($c \notin \mathcal{V}$). This targeted sampling of negative contexts will ensure dissimilar embeddings of disease words ($w \in \mathcal{V}$) and non-disease contexts ($c \notin \mathcal{V}$), thus scattering them in the vector space. However, sampling negative examples only from the set of non-disease contexts may lead to overfitting and thus, we introduce a sampling parameter π_s which controls the probability of drawing a *negative* example from non-disease contexts ($c \in \mathcal{V}$) versus disease contexts ($c \in \mathcal{V}$). *Dis2Vec*'s objective for $(w, c) \in \mathcal{D}_{(d)}$ is shown below in equation 4.

$$\begin{aligned} l_{\mathcal{D}_{(d)}} = \sum_{(w,c)\in\mathcal{D}_{(d)}} \Big(& \log \sigma(\mathbf{w}\cdot\mathbf{c}) \\ & + k \cdot [P(x_k < \pi_s)\mathbb{E}_{c_N \sim P_{D_{c\notin\mathcal{V}}}}[\log \sigma(-\mathbf{w}\cdot\mathbf{c_N})] \\ & + P(x_k \geq \pi_s)\mathbb{E}_{c_N \sim P_{D_{c\in\mathcal{V}}}}[\log \sigma(-\mathbf{w}\cdot\mathbf{c_N})]] \Big) \end{aligned} \tag{4}$$

where $x_k \sim U(0, 1)$, U(0,1) being the uniform distribution on the interval [0,1]. If $x_k < \pi_s$, we sample a negative context c_N from the unigram distribution $P_{D_{c\notin\mathcal{V}}}$ where $D_{c\notin\mathcal{V}}$ is the collection of (w, c) pairs for which $c \notin \mathcal{V}$ and $P_{D_{c\notin\mathcal{V}}} = \frac{\#(c)^\alpha}{\sum_{c\notin\mathcal{V}} \#(c)^\alpha}$ where α is the smoothing parameter. For values of $x_k \geq \pi_s$, we sample c_N from the unigram distribution $P_{D_{c\in\mathcal{V}}}$ and $P_{D_{c\in\mathcal{V}}} = \frac{\#(c)^\alpha}{\sum_{c\in V} \#(c)^\alpha}$. Therefore, optimizing the objective in equation 4 will have a tendency to generate disease specific word embeddings for values of $\pi_s \geq 0.5$ due to the reason that higher number of negative contexts (c_N) will be sampled from the set of non-disease contexts ($c \notin \mathcal{V}$) with $\pi_s \geq 0.5$.

3.3.2 Out-of-vocabulary Objective Regularization

The second category ($\mathcal{D}_{(\neg d)}$) of (w, c) pairs consists of those pairs for which both w and c are not in $\mathcal{V}$ ($w \notin \mathcal{V} \wedge c \notin \mathcal{V}$). These pairs are uninformative to us in generating disease specific word embeddings since both w and c are not a part of $\mathcal{V}$. However, minimizing the dot products, i.e. optimizing the objective $\log \sigma(-\mathbf{w}\cdot\mathbf{c})$ for these non-disease word-context pairs will scatter them in the embedding space (dissimilar embeddings) and thus, a word $w \notin \mathcal{V}$ can have similar embeddings (or, get closer) to a word $w \in \mathcal{V}$ which should be avoidable in our scenario. Therefore, we need to maximize $\log \sigma(\mathbf{w}\cdot\mathbf{c})$ for these (w, c) pairs in order to achieve similar (or, closer) embeddings. We adopt the basic objective function of *SGNS* for $(w, c) \in \mathcal{D}_{(\neg d)}$ as shown below in equation 5.

$$l_{\mathcal{D}_{(\neg d)}} = \sum_{(w,c)\in\mathcal{D}_{(\neg d)}} \Big(\log \sigma(\mathbf{w}\cdot\mathbf{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\mathbf{w}\cdot\mathbf{c_N})] \Big) \tag{5}$$

3.3.3 Vocabulary Driven Objective Minimization

Lastly, the third category ($\mathcal{D}_{(d)(\neg d)}$) consists of (w, c) pairs where either w is in $\mathcal{V}$ or c is in $\mathcal{V}$ ($w \in \mathcal{V} \oplus c \in \mathcal{V}$) but both cannot be in $\mathcal{V}$. Consider an arbitrary (w, c) pair belonging to $\mathcal{D}_{(d)(\neg d)}$. As per the objective (equation 3) of *SGNS*, two words are similar to each other if they share the same contexts or if they tend to appear in the contexts of each other (and preferably both). If $w \in \mathcal{V}$ and $c \notin \mathcal{V}$, then maximizing $\log \sigma(\mathbf{w}\cdot\mathbf{c})$ will have the tendency to generate similar embeddings for the disease word $w \in \mathcal{V}$ and non-disease words $\notin \mathcal{V}$ which share the same non-disease context $c \notin \mathcal{V}$. On the other word, if $c \in \mathcal{V}$ and $w \notin \mathcal{V}$, then maximizing $\log \sigma(\mathbf{w}\cdot\mathbf{c})$ will drive the embedding of the non-disease word $w \notin \mathcal{V}$ closer to the embeddings of disease words $\in \mathcal{V}$ sharing the same disease context $c \in \mathcal{V}$. Therefore, we posit that the dot products for this category of (w, c) pairs should be minimized, i.e. the objective $\log \sigma(-\mathbf{w}\cdot\mathbf{c})$ should be optimized in order to ensure dissimilar embeddings for these (w, c) pairs. However, minimizing the dot products of all such word-context pairs may lead to over-penalization

and thus we introduce an objective selection parameter π_o which controls the probability of selecting $\log\sigma(-\mathbf{w}\cdot\mathbf{c})$ versus $\log\sigma(\mathbf{w}\cdot\mathbf{c})$. The objective for $(w,c)\in\mathcal{D}_{(d)(\neg d)}$ is shown below in equation 6.

$$l_{\mathcal{D}_{(d)(\neg d)}} = \sum_{(w,c)\in\mathcal{D}_{(d)(\neg d)}} \Big(P(z<\pi_o)\log\sigma(-\mathbf{w}\cdot\mathbf{c}) + P(z\geq\pi_o)\log\sigma(\mathbf{w}\cdot\mathbf{c}) \Big) \quad (6)$$

where $z\sim U(0,1)$, U(0,1) being the uniform distribution over the interval [0,1]. If $z<\pi_o$, $\log\sigma(-\mathbf{w}\cdot\mathbf{c})$ gets optimized, otherwise *Dis2Vec* optimizes $\log\sigma(\mathbf{w}\cdot\mathbf{c})$. Therefore, optimizing the objective in equation 6 will have a tendency to generate disease specific embeddings with values of $\pi_o\geq 0.5$ due to the reason that the objective $\log\sigma(-\mathbf{w}\cdot\mathbf{c})$ will be selected for optimization with a higher probability over $\log\sigma(\mathbf{w}\cdot\mathbf{c})$.
Finally, the overall objective of *Dis2Vec* comprising all three categories of (w,c) pairs can be defined as below.

$$l_{Dis2Vec} = l_{\mathcal{D}_{(d)}} + l_{\mathcal{D}_{(\neg d)}} + l_{\mathcal{D}_{(d)(\neg d)}} \quad (7)$$

Similar to *SGNS*, the objective in equation 7 is trained in an online fashion using stochastic gradient updates over the three categories of (w,c) pairs.

Algorithm 1: *Dis2Vec* model

Input : Unstructured corpus $\mathcal{D}=\{(w,c)\}$, $\mathcal{V}$
Output: word embeddings $\mathbf{w}\forall w\in\mathcal{W}$, column embeddings $\mathbf{c}\forall c\in\mathcal{C}$

Categorize $\mathcal{D}$ into 3 types: $\mathcal{D}_{(d)} = \{(w,c): w\in\mathcal{V}\wedge c\in\mathcal{V}\}$, $\mathcal{D}_{(\neg d)} = \{(w,c): w\notin\mathcal{V}\wedge c\notin\mathcal{V}\}$, $\mathcal{D}_{(d)(\neg d)} = \{(w,c): w\in\mathcal{V}\oplus c\in\mathcal{V}\}$
for *each* $(w,c)\in\mathcal{D}$ **do**
  if $(w,c)\in\mathcal{D}_{(d)}$ **then**
    train the (w,c) pair using the objective in equation 4
  else if $(w,c)\in\mathcal{D}_{(\neg d)}$ **then**
    train the (w,c) pair using the objective in equation 5
  else
    train the (w,c) pair using the objective in equation 6

3.4 Parameters in *Dis2Vec*

Dis2Vec inherits all the parameters of *SGNS*, such as dimensionality (T) of the word embeddings, window size (L), number of negative samples (k) and context distribution smoothing (α). It also introduces two new parameters - the objective selection parameter (π_o) and the sampling parameter (π_s). The explored values for each of the aforementioned parameters are shown in Table 6.

4. EXPERIMENTAL EVALUATION

We evaluated *Dis2Vec* against several state-of-the art methods. In this section, we first provide a brief description of our experimental setup, including the disease news corpus, human annotated taxonomy and the domain information used as the vocabulary $\mathcal{V}$ for the process. We present our experimental findings in Section 4.2 where we have compared our model against several baselines and also explore its applicability to emerging diseases.

4.1 Experimental Setup

4.1.1 Corpus

We collected a dataset corresponding to a corpus of public health-related news articles in English extracted from HealthMap [7], a prominent online aggregator of news articles from all over the world for disease outbreak monitoring and real-time surveillance of emerging public health threats. Each article contains the following information - textual content, disease tag, reported date and location information in the form of (lat, long) coordinates. The articles were reported during the time period 2010 to 2014 and correspond to locations from all over the world. The textual content of each article was pre-processed by sentence splitting, tokenization and lemmatization via BASIS Technologies' Rosette Language Processing (RLP) tools [17]. After pre-processing, the corpus consisting of 124850 articles was found to contain 1607921 sentences, spanning 52679298 words. Words that appeared less than 5 times in the corpus were ignored, resulting in a vocabulary of 91178 words.

4.1.2 Human Annotated Taxonomy

Literature reviews were conducted for each of the 39 infectious diseases of interest in order to make classifications for transmission methods, transmission agents, clinical symptoms and exposures or risk factors. These 39 diseases were selected such that no bias is included in the process, i.e. they represent a diversity of infectious diseases ranging from emerging (*H7N9*, *Ebola*) to endemic (*dengue*, *avian influenza*) to rare (*plague*, *hantavirus*, *yellow fever*).

Methods of transmission were first classified into 8 subcategories - *direct contact*, *droplet*, *airborne*, *zoonotic*, *vectorborne*, *waterborne*, *foodborne*, and *environmental*. For many diseases, multiple subcategories of transmission methods could be assigned. Transmission agents were classified into 8 subcategories - *wild animal*, *fomite*, *fly*, *mosquito*, *bushmeat*, *flea*, *tick* and *domestic animal*. The category of clinical symptoms was broken down into 8 subcategories: *general*, *gastrointestinal*, *respiratory*, *nervous system*, *cutaneous*, *circulatory*, *musculoskeletal*, and *urogenital*. A full list of the symptoms within each subcategory can be found in Table 2. For disease exposures or risk factors, 7 subcategories were assigned based on those listed/most commonly reported in the literature. The subcategories include: *healthcare facility* , *healthcare worker*, *schoolchild*, *mass gathering*, *travel*, *animal exposure*, and *weakened immune system*. The *animal exposure* category was further broken down into *farmer*, *veterinarian*, *market* and *slaughter*. For some diseases, there were no risk factors listed, and for other diseases, multiple exposures were assigned.

Table 2: Symptom categories and corresponding words.

Symptom Category	Words
General	Fever, chill, weight loss, fatigue, lethargy, headache
Gastrointestinal	Abdominal pain, nausea, diarrhea, vomiting
Respiratory	Cough, runny nose, sneezing, chest pain, sore throat, pneumonia, dyspnea
Nervous system	Mental status, paralysis, paresthesia, encephalitis, meningitis
Cutaneous	Rash, sore, pink eye
Circulatory	Hemorrhagic
Musculoskeletal	Joint pain, muscle pain, muscle ache

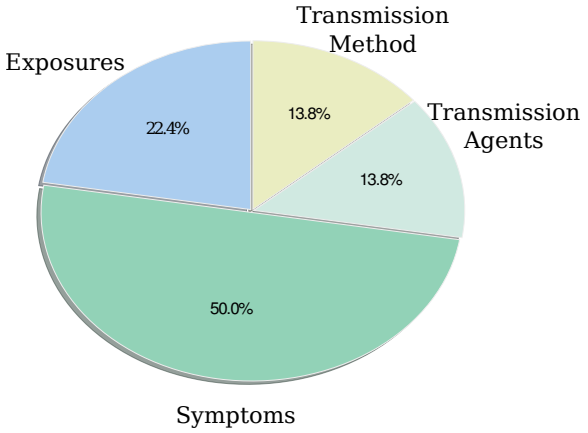


Figure 3: Distribution of word counts corresponding to each taxonomical category in the disease vocabulary ($\mathcal{V}$). Words related to clinical symptoms constitute the majority of $\mathcal{V}$ with relatively much smaller percentages of terms related to exposures, transmission agents and transmission methods

4.1.3 Disease Vocabulary $\mathcal{V}$

Disease vocabulary $\mathcal{V}$ is provided as prior knowledge to *Dis2Vec* in order to generate disease specific word embeddings as explained in section 3.3. $\mathcal{V}$ is represented by a flat list of disease-related terms consisting of disease names (*influenza*, *h7n9*, *plague*, *ebola*, etc.), all possible words related to transmission methods(*vectorborne*, *foodborne*, *waterborne*, etc.), all possible words related to transmission agents (*flea*, *domestic animal*, *mosquito*, etc.), all possible words related to clinical symptoms (*fever*, *nausea*, *paralysis*, *cough*, *headache*, etc.) and all possible words related to exposures or risk factors (*healthcare facility*, *slaughter*, *farmer*, etc.). We treat the multi-word expressions (e.g. *healthcare facility*, *sore throat*) in $\mathcal{V}$ as phrases, i.e. we learn a single embedding for these expressions, not a composite embedding of its individual terms. Total number of words in $\mathcal{V}$ is found to be 103. In Figure 3, we show the distribution of word counts associated with different taxonomical categories in the disease vocabulary ($\mathcal{V}$). As depicted in Figure 3, half of the words in $\mathcal{V}$ are terms related to clinical symptoms followed by exposures or risk factors (22.4%), transmission methods (13.8%) and transmission agent(s) (13.8%).

4.1.4 Baselines

We compared the following baseline models with *Dis2Vec* on the four disease characterization tasks.

• ***SGNS***: Unsupervised skip-gram model with negative sampling [13] described in section 3.2.

• ***SGHS***: skip-gram model trained using the hierarchical softmax algorithm [13] instead of negative sampling.

• ***CBOW***: Continuous bag-of-words model described in [12]. Unlike skip-gram models, the training objective of the *CBOW* model is to correctly predict the target word given its contexts (surrounding words). *CBOW* is denoted as a bag-of-words model as the order of words in the contexts does not have any impact on the model.

All models (both baselines and *Dis2Vec*) were trained on the HealthMap corpus using a T-dimensional word embedding via gensim's word2vec software [18]. We explored a large space of parameters for each model. In Table 6, we provide the list of parameters, the explored values for each parameter and the applicable models corresponding to each parameter. Apart from the parameters listed in Table 6, we also applied the sub-sampling technique developed by Mikolov et al. [13] to each model in order to counter the imbalance between common words (such as, *is*, *of*, *the*, *a*, etc.) and rare words. In the context of NLP, these common words are referred to as *stop words*. For more details on the sub-sampling techniques, please see Mikolov et al. [13]. Our initial experiments (not reported) demonstrated that both the baselines and *Dis2Vec* showed improved results on the disease characterization tasks with sub-sampling versus without sub-sampling.

4.1.5 Accuracy Metric

We evaluate the automatic taxonomy generation methods such that for a taxonomical characteristic of a disease, models that generate similar set of terms (top words) as the human annotated ones are more preferable. As such, we use cosine similarity in a min-max setting between the aforementioned sets for a particular characterization category as our accuracy metric. The overall accuracy of a model for a category can be found by averaging the accuracy values across all diseases of interest. This is a bounded metric (between 0 and 1) where higher values indicate better model performance. We can formalize the metric as follows. Let D be the disease and C be the taxonomical category under investigation. Furthermore, let $C_1, C_2, \cdots, C_N$ be all possible terms or words related to C and $H_1, H_2, \cdots, H_M$ be the human annotated words. Then the characterization accuracy corresponding to category C and disease D is given below in equation 8.

$$Accuracy(C, D) = \frac{1}{M}\sum_{j=1}^{M}\frac{cosine(\mathbf{D}, \mathbf{H_j}) - \min_i cosine(\mathbf{D}, \mathbf{C_i})}{\max_i cosine(\mathbf{D}, \mathbf{C_i}) - \min_i cosine(\mathbf{D}, \mathbf{C_i})} \quad (8)$$

where $\mathbf{D}$, $\mathbf{H_j}$ and $\mathbf{C_i}$ represent the word embeddings for D, H_j and C_i. $\min_i cosine(\mathbf{D}, \mathbf{C_i})$ and $\max_i cosine(\mathbf{D}, \mathbf{C_i})$ represent the maximum and minimum cosine similarity values between $\mathbf{D}$ and the word embeddings of the terms related to C. Therefore, equation 8 indicates that if the human annotated word $\mathbf{H_j}$ is among the top words found by the word2vec model for the category C, then the ratio in the numerator is high leading to high accuracy and vice versa.

4.2 Results

In this section we try to ascertain the efficacy and the applicability of *Dis2Vec* by investigating some of the pertinent questions related to the problem of disease characterization.

Sample-vs-objective: which is the better method to incorporate disease vocabulary information into *Dis2Vec*? As described in Section 3, there are primarily two different ways by which disease vocabulary information ($\mathcal{V}$) guides the generation of embeddings for *Dis2Vec* (a) by modulating negative sampling parameter (π_s) for disease word-context pairs ($(w, c) \in \mathcal{D}_{(d)}$) referred to as *Dis2Vec-sample* and (b) by modulating the objective selection parameter (π_o) for non-disease words or non-disease contexts ($(w, c) \in \mathcal{D}_{(d)(\neg d)}$) referred to as *Dis2Vec-objective*. We investigate the importance of these two strategies by comparing the accuracies for each strategy individually (*Dis2Vec-sample* and *Dis2Vec-objective*) as well as combined together (*Dis2Vec-combined*) under the best parameter configuration for a particular task in Table 3. As can be seen, no single strategy is best across all tasks. Henceforth, we select the

best performing strategy for a particular task as our *Dis2Vec* in the next Table 4.

Table 3: Comparative performance evaluation of *Dis2Vec-combined* against *Dis2Vec-objective* and *Dis2Vec-sample* across the 4 characterization tasks under the best parameter configuration for that model and task combination. The value in each cell represents the overall accuracy across the 39 diseases for that particular model and characterization task. We use equation 8 as the accuracy metric in this table.

Characterization tasks	*Dis2Vec-sample*	*Dis2Vec-objective*	*Dis2Vec-combined*
Symptoms	0.635	**0.945**	0.940
Exposures	0.590	0.540	**0.597**
Transmission methods	**0.794**	0.754	0.734
Transmission agents	0.505	0.506	**0.516**
Overall average accuracy	0.631	0.686	**0.697**

Table 4: Comparative performance evaluation of *Dis2Vec* against *SGNS*, *SGHS* and *CBOW* across the 4 characterization tasks under the best parameter configuration for that model and task combination. The value in each cell represents the overall accuracy across the 39 diseases for that particular model and characterization task. We use equation 8 as the accuracy metric in this table.

Characterization tasks	*CBOW*	*SGHS*	*SGNS*	*Dis2Vec*
Symptoms	0.498	0.560	0.620	**0.945**
Exposures	0.383	0.498	**0.605**	0.597
Transmission methods	0.481	0.765	0.792	**0.794**
Transmission agents	0.274	0.466	0.498	**0.516**
Overall average accuracy	0.409	0.572	0.629	**0.713**

Does disease vocabulary information improve disease characterization? *Dis2Vec* was designed to incorporate disease vocabulary information in order to guide the generation of disease specific word embeddings. To evaluate the importance of such vocabulary information in *Dis2Vec*, we compare the performance of *Dis2Vec* against the baseline word2vec models described in section 4.1.4 under the best parameter configuration for a particular task. These baseline models do not permit incorporation of any vocabulary information due to their unsupervised nature. Table 4 presents the accuracy of the models for the 4 disease characterization tasks - symptoms, exposures, transmission methods and transmission agents. As can be seen, *Dis2Vec* performs the best for 3 tasks and in average. It is also interesting to note that *Dis2Vec* achieves higher performance gain over the baseline models for the symptoms category than the other categories. The superior performance of *Dis2Vec* in the symptoms category can be attributed to two factors - (a) higher percentage of symptom words in the disease vocabulary $\mathcal{V}$ (see Figure 3) and (b) higher occurrences of symptom words in the HealthMap news corpus. News articles reporting a disease outbreak generally tend to focus more on the symptoms related to the disease rather than the other categories. Given the functionality of Dis2Vec, higher occurrences of symptom terms in outbreak news reports will lead to generation of efficient word embeddings for characterizing disease symptoms.

What are beneficial parameter configurations for characterizing diseases? To identify which parameter settings are beneficial for characterizing diseases, we looked at the best parameter configuration of all the 6 models on each task. We then counted the number of times each parameter setting was chosen in these configurations (see Table 6). We compared standard settings of each parameter as explored in previous research [11]. For the new parameters π_s and π_o introduced by *Dis2Vec*, we chose the values 0.3, 0.5 and 0.7 in order to analyze the impact of these parameters with values < 0.5 and ≥ 0.5. For *Dis2Vec-objective* and *Dis2Vec-combined*, some trends emerge regarding the parameter π_o that these two models consistently benefit from values of $\pi_o \geq 0.5$ validating our claims in section 3.3 that when $\pi_o \geq 0.5$, disease words and non-disease words get scattered from each other in the vector space, thus tending to generate disease specific embeddings. However, for π_s we observe mixed trends. As expected, *Dis2Vec-sample* benefits from higher values of sampling parameter $\pi_s \geq 0.5$. But *Dis2Vec-combined* seems to prefer lower values of $\pi_s < 0.5$ and higher values of $\pi_o \geq 0.5$ for the disease characterization tasks. For the smoothing parameter(α), all the applicable models prefer smoothed unigram distribution ($\alpha = 0.75$) for negative sampling except *Dis2Vec-combined* which is in favor of unsmoothed distribution ($\alpha = 1.0$) for characterizing diseases. For the number of *negative* samples k, all the applicable models seem to benefit from $k > 1$ except *Dis2Vec-combined* which seems to prefer $k = 1$. For the window size (L), all the models prefer smaller-sized context windows (either 5 or 10) except *SGHS* which prefers larger-sized windows ($L > 10$) for characterizing diseases. Finally, regarding the dimensionality (T) of the embeddings, *Dis2Vec-combined*, *Dis2Vec-sample* and *SGNS* are in equal favor of both 300 and 600 dimensions. *Dis2Vec-objective* and *SGHS* prefer 300 dimensions and *CBOW* is in favor of 600 dimensions for characterizing diseases.

Table 7: Comparative performance evaluation of *Dis2Vec* against *SGNS*, *SGHS* and *CBOW* across the 4 characterization tasks for each class of diseases (emerging, endemic and rare) under the best parameter configuration for a particular {disease class, task, model} combination. We use equation 8 as the accuracy metric in this table.

Class	Tasks	*CBOW*	*SGHS*	*SGNS*	*Dis2Vec*
Emerging	Symptoms	0.589	0.671	0.722	**0.977**
	Exposures	0.356	0.495	0.516	**0.679**
	Transmission methods	0.407	0.885	0.898	**0.945**
	Transmission agents	0.528	0.587	0.795	**0.975**
Endemic	Symptoms	0.453	0.583	0.671	**0.930**
	Exposures	0.421	0.512	**0.642**	0.631
	Transmission methods	0.472	0.820	0.851	**0.856**
	Transmission agents	0.164	0.399	0.408	**0.415**
Rare	Symptoms	0.506	0.536	0.599	**0.949**
	Exposures	0.377	0.525	0.616	**0.670**
	Transmission methods	0.503	0.760	0.755	**0.775**
	Transmission agents	0.320	**0.522**	0.512	0.515

Importance of taxonomical categories - how should we construct the disease vocabulary? We followup our previous analysis by investigating the importance of words related to each taxonomical category in constructing the disease vocabulary towards final characterization accuracy. To evaluate a particular category, we used a truncated disease vocabulary consisting of disease names and the words in the corresponding category to drive the discovery of word embeddings in *Dis2Vec* under the best parameter configuration for that category. We compared the accuracy of each of these conditions (*Dis2Vec* (exposures), *Dis2Vec* (transmission methods), *Dis2Vec* (transmission agents), *Dis2Vec* (symptoms)) against *Dis2Vec* (full vocabulary) across the 4 characterization tasks. Table 5 presents our results for this analysis and provides multiple insights as follows. (a) Constructing the vocabulary with words related to all the categories leads to better characterization across all the tasks. (b) As expected, *Dis2Vec* (symptoms) is the second best performing model for the symptoms category but it's per-

Table 5: Comparative performance evaluation of *Dis2Vec* with full vocabulary against each of the 6 conditions of *Dis2Vec* with a truncated vocabulary across the 4 characterization tasks where the truncated vocabulary consists of disease names and all possible terms related to a particular taxonomical category. We use equation 8 as the accuracy metric in this table.

Characterization tasks	*Dis2Vec* (Exposures)	*Dis2Vec* (Transmission methods)	*Dis2Vec* (Transmission agents)	*Dis2Vec* (Symptoms)	*Dis2Vec* (full vocabulary)
Symptoms	0.597	0.581	0.165	0.883	**0.945**
Exposures	0.554	0.557	0.315	0.416	**0.597**
Transmission methods	0.748	0.768	0.517	0.455	**0.794**
Transmission agents	0.446	0.459	0.467	0.457	**0.516**

Table 6: Comparison of different parameter settings for each model, measured by the number of characterization tasks in which the best configuration had that parameter setting. Non-applicable combinations are marked by 'NA'

Method	T 300 : 600	L 5 : 10 : 15	k 1 : 5 : 15	α 0.75 : 1	π_s 0.3 : 0.5 : 0.7	π_o 0.3 : 0.5 : 0.7
Dis2Vec-combined	2 : 2	3 : 1 : 0	2 : 1 : 1	1 : 3	4 : 0 : 0	0 : 2 : 2
Dis2Vec-sample	2 : 2	2 : 1 : 1	1 : 1 : 2	4 : 0	1 : 2 : 1	NA
Dis2Vec-objective	3 : 1	2 : 2 : 0	1 : 1 : 2	3 : 1	NA	2 : 0 : 2
SGNS	2 : 2	2 : 2 : 0	0 : 2 : 2	2 : 2	NA	NA
SGHS	3 : 1	1 : 0 : 3	NA	NA	NA	NA
CBOW	0 : 4	0 : 4 : 0	NA	NA	NA	NA

formance is degraded for other tasks. The same goes for *Dis2Vec* (transmission methods) and *Dis2Vec* (transmission agents). (c) Therefore, it indicates that in order to achieve reasonable characterization accuracy for a category, we need to supply at least the words related to that category along with the disease names in constructing the vocabulary.

Can *Dis2Vec* be applied to characterize emerging, endemic and rare diseases? We classified the 39 diseases of interest into 3 classes as follows. For classifying each disease, we plotted the time series of the counts of HealthMap articles with disease tag equal to the corresponding disease. (a) **Endemic**: We considered a disease as endemic if the counts of articles were consistently high for all years with repeating shapes. E.g.- rabies, avian influenza, west nile virus. (b) **Emerging**: We considered a disease as emerging if the counts of articles were historically low, but have peaked in recent years. E.g.- Ebola, H7N9, MERS. (c) **Rare**: We considered a disease as rare if the counts were consistently low for all years with or without sudden spikes. E.g.- plague, chagas, japanese encephalitis. We also considered a disease as rare if the counts of articles were high in 2010/2011, but have since fallen down and depicted consistently low counts. E.g.- tuberculosis. Following classification, the distribution of emerging, endemic and rare diseases is 4 : 12 : 23 respectively. In Table 7, we compared the accuracy of *Dis2Vec* against the baseline word2vec models for each class of diseases across the 4 characterization tasks under the best parameter configuration for a particular {disease class, task, model} combination. It can be seen that *Dis2Vec* is the best performing model for majority of the {disease class, task} combinations except {endemic, exposures} and {rare, transmission agents}. It is interesting to note that for the symptoms category, *Dis2Vec* performs better than the baseline models across all the disease classes. Irrespective of disease class, news reports generally mention the symptoms of the disease while reporting an outbreak. As the characteristics of the emerging diseases are relatively unknown w.r.t. endemic and rare, news media reports also tend to focus on other categories (exposures, transmission methods, transmission agents) apart from the symptoms to create awareness among the general public. Therefore, *Dis2Vec* and the baselines perform better overall for the emerging diseases in comparison to endemic and rare diseases. However, *Dis2Vec* outperforms the baselines for characterizing symptoms and exposures of emerging diseases. For endemic and rare diseases, *Dis2Vec* achieves higher accuracy than the baseline models w.r.t. the symptoms category. For other categories, *Dis2Vec* performs better overall, although the performance gain is not high in comparison to the symptoms. It is to be noted that *Dis2Vec* achieves reasonable accuracy for characterizing rare diseases even though the number of articles related to these diseases is very few in HealthMap corpus leading to under-represented categories. In Figure 4, we show the top words selected for each category of an emerging disease (H7N9), an endemic disease (avian influenza) and a rare disease (plague) across all the models. The human annotated words corresponding to each category of these diseases can be found in Table 1. We selected these 3 diseases due to their public health significance and the fact that these diseases have complete coverage across all the taxonomical categories (see Table 1). It is interesting to note that for H7N9, the top words found by *Dis2Vec* for the symptoms category contain all the human annotated words *fever*, *cough* and *pneumonia*, while the top words found by *SGNS* only contain the word *fever*. For exposures (H7N9), *Dis2Vec* is able to capture three human annotated words *animal exposure*, farmer, slaughter. However, *SGNS* is only able to capture the word *animal*. For the symptoms category of the rare disease plague, *Dis2Vec* is able to detect three human annotated words *sore*, *fever* and *headache* with *SGNS* only being able to detect the word *fever*. Moreover, *Dis2Vec* is able to characterize the transmission method of plague as *vectorborne* with *SGNS* failing to do so.

5. CONCLUSIONS

Classical word2vec methods such as *SGNS* and *SGHS* have been applied to solve a variety of linguistic tasks with considerable accuracy. However, such methods fail to generate satisfactory embeddings for highly specific domains such as healthcare where uncovering the relationships with respect to domain specific words is of greater importance than the non-domain ones. These algorithms are by design unsupervised and do not permit the inclusion of domain information to find interesting embeddings. In this paper, we have proposed *Dis2Vec*, a disease specific word2vec framework that given an unstructured news corpus and domain knowledge in terms of important words, can find interesting disease characterizations. We demonstrated the strength of

our model by comparing it against three classical word2vec methods on four disease characterization tasks. *Dis2Vec* exhibits the best overall accuracy for 3 tasks across all the diseases and in general, its relative performance improvement is found to be empirically dependent on the amount of supplied domain knowledge. Consequently, *Dis2Vec* works especially well for characteristics with more domain knowledge (symptoms) and is found to be a promising tool to analyze different class of diseases viz. emerging, endemic and rare. In future, we aim to analyze a greater variety of diseases and try to ascertain common relationships between such diseases across different geographical regions.

Supplementary Information

https://github.com/sauravcsvt/Dis2Vec_supplementary.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

6. REFERENCES

[1] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the ACL*, pages 238–247, 2014.

[2] M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[3] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[4] P. Chakraborty, N. Ramakrishnan, et al. Forecasting a moving target: Ensemble models for ILI case count predictions. In *Proceedings of the SIAM International Conference on Data Mining*, pages 262–270, 2014.

[5] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 160–167. ACM, 2008.

[6] R. Collobert, J. Weston, et al. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[7] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, 2008.

[8] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the ACL*, pages 302–308, 2014.

[9] O. Levy and Y. Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on CoNLL*, pages 171–180, 2014.

[10] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *27th Annual Conference on Neural Information Processing Systems*, pages 2177–2185, 2014.

[11] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225, 2015.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *26th Annual Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.

[14] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the NAACL*, pages 746–751, 2013.

[15] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

[16] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[17] N. Ramakrishnan, P. Butler, S. Muthiah, et al. 'beating the news' with embers: Forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD*, pages 1799–1808, New York, NY, USA, 2014. ACM.

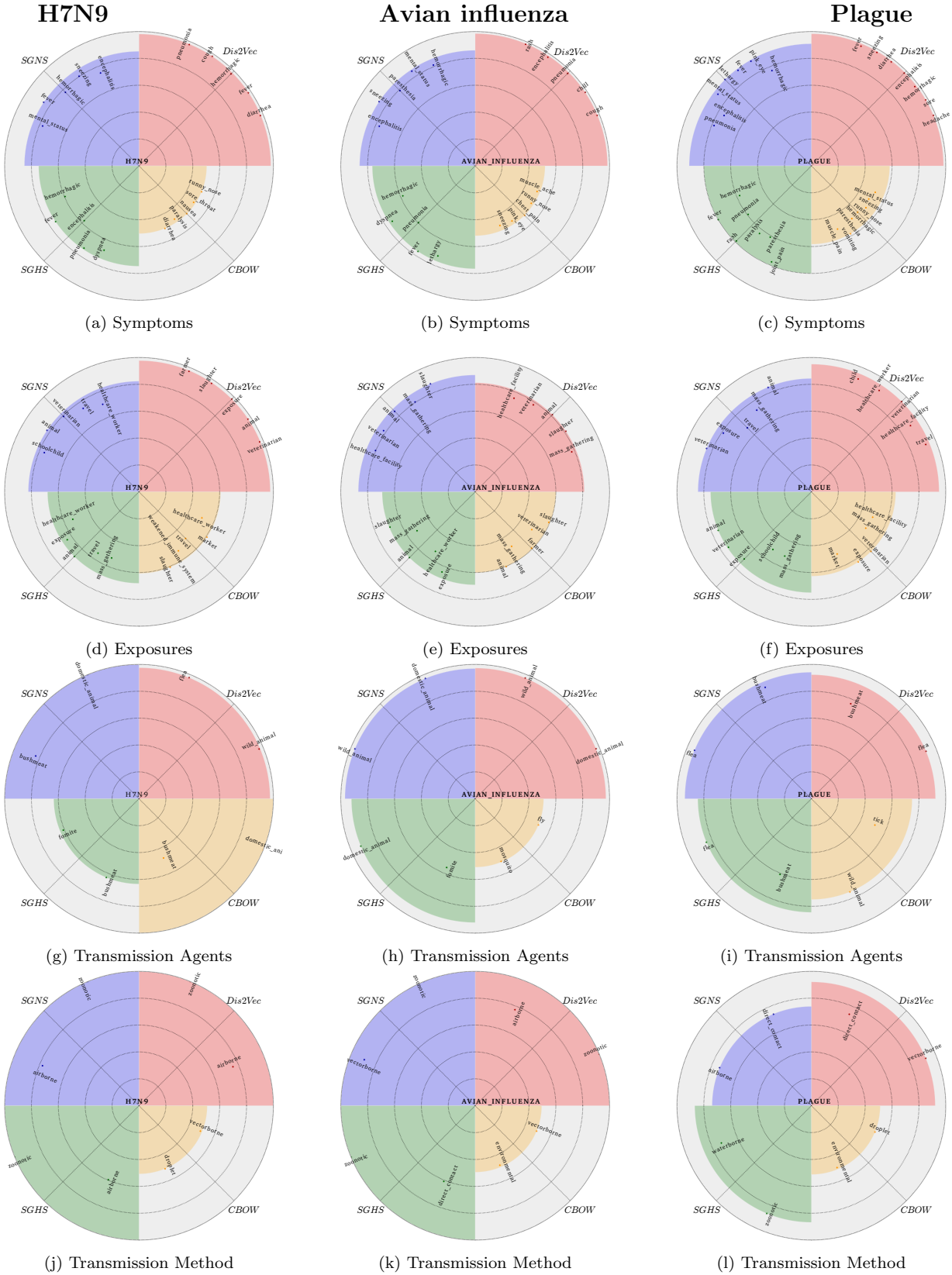
[18] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[19] T. Rekatsinas, S. Ghosh, N. Ramakrishnan, et al. Sourceseer: Forecasting rare disease outbreaks using multiple data sources. In *Proceedings of the SIAM International Conference on Data Mining*, pages 379–387, 2015.

[20] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the ACL*, pages 384–394, 2010.

[21] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[22] Z. Wang, P. Chakraborty, N. Ramakrishnan, et al. Dynamic poisson autoregression for influenza-like-illness case count prediction. In *Proceedings of the 21th ACM SIGKDD*, pages 1285–1294. ACM, 08 2015.



(a) Symptoms (b) Symptoms (c) Symptoms

(d) Exposures (e) Exposures (f) Exposures

(g) Transmission Agents (h) Transmission Agents (i) Transmission Agents

(j) Transmission Method (k) Transmission Method (l) Transmission Method

Figure 4: Case study for emerging, endemic and rare diseases: Disease characterization accuracy plot for *Dis2Vec* (first quadrant, red), *SGNS* (second quadrant, blue), *SGHS* (third quadrant, green), and *CBOW* (fourth quadrant, orange) w.r.t. H7N9 (left, emerging), avian influenza (middle, endemic) and plague (right, rare). The shaded area in a quadrant indicates the cosine similarity (scaled between 0 and 1) of the top words found for the category of interest using corresponding model, as evaluated against the human annotated words (see Table 1). The top words found for each model is shown in the corresponding quadrant with radius equal to its average similarity with the human annotated words for the disease. *Dis2Vec* shows best overall performance with noticeable improvements for symptoms w.r.t. all diseases.