Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data

Bin Zou
Department of Computer Science
University College London
United Kingdom
bin.zou.14@ucl.ac.uk

Vasileios Lampos
Department of Computer Science
University College London
United Kingdom
v.lampos@ucl.ac.uk

Ingemar J. Cox*
Department of Computer Science
University College London
United Kingdom
i.cox@ucl.ac.uk

ABSTRACT

A considerable body of research has demonstrated that online search data can be used to complement current syndromic surveillance systems. The vast majority of previous work proposes solutions that are based on supervised learning paradigms, in which historical disease rates are required for training a model. However, for many geographical regions this information is either sparse or not available due to a poor health infrastructure. It is these regions that have the most to benefit from inferring population health statistics from online user search activity. To address this issue, we propose a statistical framework in which we first learn a supervised model for a region with adequate historical disease rates, and then transfer it to a target region, where no syndromic surveillance data exists. This transfer learning solution consists of three steps: (i) learn a regularized regression model for a source country, (ii) map the source queries to target ones using semantic and temporal similarity metrics, and (iii) re-adjust the weights of the target queries. It is evaluated on the task of estimating influenza-like illness (ILI) rates. We learn a source model for the United States, and subsequently transfer it to three other countries, namely France, Spain and Australia. Overall, the transferred (unsupervised) models achieve strong performance in terms of Pearson correlation with the ground truth (> .92 on average), and their mean absolute error does not deviate greatly from a fully supervised baseline.

ACM Reference Format:

Bin Zou, Vasileios Lampos, and Ingemar J. Cox. 2019. Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313477

1 INTRODUCTION

Syndromic surveillance systems aim to provide timely estimates about the prevalence of a disease in a population. Their main source of information is based on doctor assessments about the probable health status of patients given a set of symptoms. For example, to monitor the rate of influenza, syndromic surveillance relies on a network of doctors who report on a daily or weekly basis the number of patients exhibiting related symptoms, such as fever, cough or a sore throat. Recent research efforts have shown that this traditional approach can be complemented by alternative methods trained on data from online user activity, e.g. social media or online search behavior [43]. Applications vary from modelling dengue fever [24] to depression [17], but particular research focus has been drawn to influenza, an infectious disease that is responsible for 290-650,000 deaths worldwide on an annual basis.[1] Data from the microblogging platform of Twitter [15, 32, 50] as well as from search engines [21, 35, 53, 67] combined with statistical natural language processing methods have produced promising outcomes, which in some occasions have been incorporated into national influenza surveillance schemes [9, 63]. The main advantages of these complementary methods are timeliness, and sampling from a larger segment of the population, including people who may not visit a doctor while being ill. It is also commonly cited that such approaches may be very useful in regions where health infrastructure is poor or absent. However, this is often impractical as the proposed machine learning solutions rely on training data which apart from the user-generated inputs, need to contain confirmed disease rates at the target location, broadly referred to as "ground truth". This data is typically provided by existing syndromic surveillance systems. Hence, for locations where ground truth is not available, user-data driven approaches are not realistically applicable.

In this paper, we propose a statistical framework to circumvent problems associated with no training data in some geographic regions. Our approach is based on the broad notion of transfer learning, where we aim to transfer parts of the knowledge gained while solving a certain task to better solve a different, but related one [49]. In particular, our goal is to transfer a well-performing disease rate inference model from a source location, where supervised learning is possible, to a target location, where supervision is not possible, given the lack of ground truth. We focus our experiments on influenza (flu) and utilize Google search query statistics as our descriptive variable for aggregate, population-level, online user activity. For example, the US Centers for Disease Control and Prevention (CDC) monitor and report influenza-like illness (ILI) rates on a weekly basis, providing sufficient ground truth to learn a function that maps online search query frequencies to these rates. In our experiments we show that we can adapt this function to derive estimates of ILI rates at different locations (outside the US). Language may or may not differ between the source and target locations. Online search statistics can be obtained for these target locations, but we assume that there is no ground truth data.

*Also affiliated with the University of Copenhagen, Denmark.

[1] World Health Organization, `who.int/mediacentre/news/statements/2017/flu/`

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313477

The proposed approach comprises 3 steps. After learning a source regression model (step 1), we seek ways to map the selected source search queries to sets of queries in the target location. To derive this mapping we deploy a hybrid metric, which combines a semantic similarity with a time series correlation component (step 2). Semantic similarities are estimated using cross-lingual or monolingual word embeddings and correlations are computed using query frequencies. Finally, query weights from the source model are transferred to the identified target queries (step 3). This framework is evaluated on three transfer learning tasks, where the source model is always based in the US, and the target countries are France, Spain and Australia. While ground truth is available for all the target countries, we only use it to evaluate the performance of the transferred models. Transferred models, assessed on four flu seasons (2012 to 2016), can accurately estimate the peak of each flu season, achieving on average Pearson correlations greater than .92 and root mean squared errors comparable to the ones obtained by the corresponding fully supervised models ($\leq$ 21.6% increase in errors). Therefore, they can be considered as practical solutions for locations that lack historical ground truth data.

Main contributions. A novel, end-to-end transfer learning framework is proposed for mapping a disease model trained on online search data from a location, where ground truth is available, to a location, where ground truth is not available. Variations of this model are investigated, exploring different query mapping functions using semantic or temporal similarities or combinations of the two. In addition, we empirically show that our approach works in three case studies, two of which require a transfer to a different language (English to French or Spanish), and one that maintains the same language (English), but demands a model transfer to a different hemisphere (US to Australia).

2 DATA SETS

We use two sources of data, namely Google search query frequency statistics and ILI rates from established health organizations.

Google search query frequency statistics. Time series of weekly search query frequencies were retrieved through Google Correlate. A frequency represents the weekly search activity of a query (number of times issued) within a geographical region. It is normalized by dividing by the total number of search queries issued during that week. This normalization controls for variations in the number of searches issued each week which can be due to a variety of causes, including summer vacations, responses to news events, and a longer-term trend of increased web usage [45]. Normalized query frequencies are subsequently standardized, such that their time series have a zero mean and a standard deviation of one. This results in expressing query frequencies under the same units for different geographical regions with potentially varying population sizes and search usage patterns. We obtained weekly frequencies of search queries from September 1, 2007 to August 31, 2016 inclusive (470 weeks) for US, France, Spain, and Australia. Given that an exhaustive list of user search queries was not available to us, we extracted them by first using a set of 12 flu-related queries per country as a seed to Google Correlate and then iterating through this process (using correlated queries as new seeds). This process extracted 34,121, 29,996, 15,673 and 8,764 queries for US, France, Spain and Australia, respectively. Queries were not limited to the topic of flu, given that various other spurious queries may also correlate with the seeds.

ILI rates. We obtained weekly ILI rates for the US, France, Spain and Australia from their established syndromic surveillance systems, namely the Centers for Disease Control and Prevention (CDC), GPs Sentinelles Network (SN), Spanish Influenza Sentinel Surveillance System (SISSS), and Australian Sentinel Practices Research Network (ASPREN), respectively.[2] ILI rates represent fraction of the population that has been diagnosed with influenza-like symptoms.[3] The data spans from September 1, 2007 to August 31, 2016 inclusive, which covers approximately 9 consecutive influenza seasons. Note that for Spain, we only have ILI rates from week 40 in a year to week 20 in the following year. The prevalence of influenza outside this period is typically very low.[4] We denote the ILI rates from each syndromic surveillance system using the corresponding country code (US, FR, ES, and AU).

In our experiments we are transferring a flu model trained on US data to one of the other three countries. To provide some insight about the difficulty of the task, we have plotted the historical ILI rates for all countries in Fig. 1. ILI rates may correlate between countries, e.g. the Pearson correlation between the US and FR rates is equal to .6 ($p \approx 3 \cdot 10^{-54}$), but peaks and troughs are occurring at different times and with very different intensity. The US and AU ILI rates are negatively correlated ($-.4$, $p \approx 8 \cdot 10^{-17}$), as expected, since these countries are situated in different hemispheres and influenza is strongly seasonal. The optimal correlation we can obtain by shifting the ILI rate time series is equal to .68 (US-ES). Notably, the metric for ILI may differ in the countries we considered in this paper. Therefore, in our experiments we are working with a standardized representation of ILI (z-score).

3 METHODS

Disease rate estimation from online search data is commonly formulated as a regression task [21, 35]. The aim is to learn a function $f \colon \mathbf{X} \rightarrow \mathbf{y}$ that maps the input space of search query frequencies, $\mathbf{X} \in \mathbb{R}^{n \times s}$, to the target variable, $\mathbf{y} \in \mathbb{R}^{n}$, representing disease rates; n denotes the number of samples and s is the size of the feature space, i.e. the number of unique search queries we are considering. More specifically, $\mathbf{X}$ contains the time series of search query frequencies, and $\mathbf{y}$ represents a rate of disease diagnoses in a population (as reported by a health agency) at corresponding times. The time interval for computing the frequency of queries is often set to one week to match the frequency of syndromic surveillance reports.

Regression approaches require observations of the target variable $\mathbf{y}$ (ground truth) for training a machine learning model. This restricts the application of such techniques to areas where historical disease rates are available. We attempt to address this limitation by proposing a transfer learning methodology, that maps an existing disease model, $f \colon \mathbf{X} \rightarrow \mathbf{y}$, from a source location, where disease rates are available, to another location, where disease rates are not possible to obtain. We define the source domain as $\mathcal{D}_{\mathrm{S}} = \{(\mathbf{x}_i, y_i)\}$,

[2]Links: **CDC** (US), `cdc.gov`; **SN** (FR), `websenti.u707.jussieu.fr/sentiweb`; **SISSS** (ES), `eng.isciii.es/ISCIII`; **ASPREN** (AU), `aspren.dmac.adelaide.edu.au`
[3]ILI is defined as the presence of high fever plus cough or sore throat [11, 46].
[4]In Figs. 1 and 2, we have set the missing ILI rates for ES to zero for visualization purposes. However, we are not using these rates to train or evaluate models for ES.

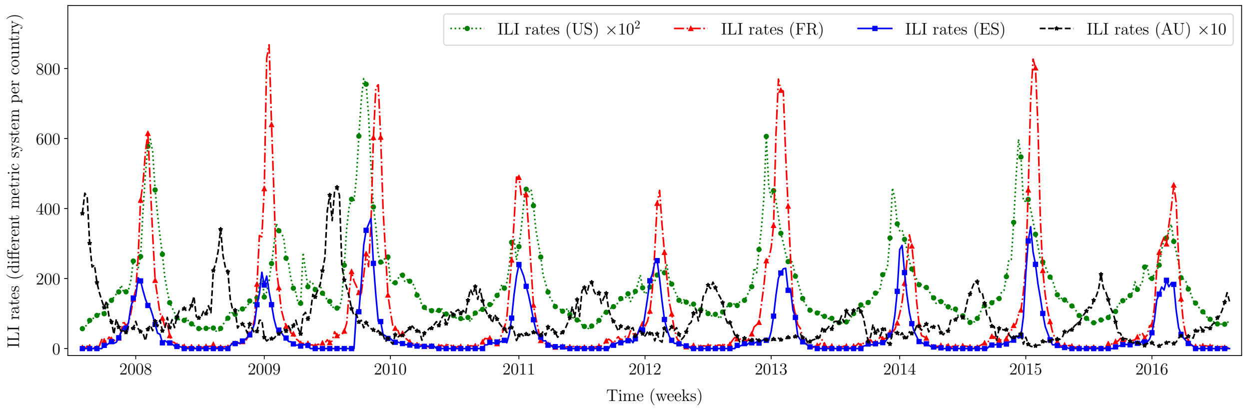

Figure 1: ILI rates for the United States (US), France (FR), Spain (ES) and Australia (AU).

$i \in \{1, \ldots, n\}$, where $\mathbf{x}_i$ is an s-dimensional vector holding the frequencies of the s queries for the time interval i, y_i is the corresponding disease rate, and n is the number of observations. The target domain is denoted by $\mathcal{D}_\mathrm{T} = \left\{\mathbf{x}'_i\right\}$, $i \in \{1, \ldots, m\}$, where $\mathbf{x}'_i$ is a t-dimensional vector of the frequencies of the t queries in the target domain that are going to be associated with the s queries in the source domain. No ground truth is available for the target domain. Note that t need not equal s, thus allowing one-to-many query mappings. In theory, the m time intervals may precede or overlap the n time intervals in the source region. In our experiments, we the m target intervals are always after the n source intervals.

3.1 User search behavior in different countries

As the transfer learning framework is detailed in the next paragraphs, it will become apparent that it is grounded on a fundamental assumption, which is that online user search behavior will be similar in the source and the target countries. Narrowing this assumption down to our specific task, this implies that the conditional probability of issuing a query q under a certain health status h (with or without experiencing disease symptoms), $P(q|h)$, will be similar for the populations of the source and the target countries. Relevant literature offers some evidence about this with regards to user search behavior for various health-related themes [1, 3, 25, 68]. In addition, we also provide some empirical evidence using our data. Table 1 shows the average query frequency over the corresponding ILI rate ratio for three basic queries in the US and AU. It also shows these ratios for translations of these queries in FR and ES (e.g. flu → grippe (FR) → gripe (ES)). The main observation is that these ratios do not vary much over the time span of our data, which is almost a decade. Although this is a limited observation, in that it does not involve many different search queries, it serves as a strong indication that user search behavior, at least for this specific area of interest, has similarities among different countries. The transfer learning framework, described in the following paragraphs, tries to exploit these similarities.

3.2 Transfer learning framework

The proposed transfer learning framework consists of three steps which are described in detail in the following sections.

3.2.1 Step 1 — Learning a regression function in the source domain. Regularized regression has been successfully applied to various text regression tasks, including the estimation of disease rates from social media or online search data [32, 35]. In this paper, we use elastic net [74] as our regression function, similarly to previous work on the topic [35, 37]. Elastic net combines ℓ_1-norm regularization, commonly known as the *lasso* [58], with ℓ_2-norm, or *ridge* [26], regularization. In addition to the sparsity encouraged by the ℓ_1-norm regularization, the ℓ_2-norm regularizer attempts to address model consistency problems that arise when collinear predictors exist in the input space [69], which is common in text regression tasks [34, 36, 54]. Given $\mathbf{X} \in \mathbb{R}^{n \times s}$ and $\mathbf{y} \in \mathbb{R}^n$ from the source domain $\mathcal{D}_\mathrm{S}$, we apply a constrained version of elastic net which solves the following optimization problem:

$$\underset{\mathbf{w}, \beta}{\operatorname{argmin}} \left(\|\mathbf{y} - \mathbf{X}\mathbf{w} - \beta\|_2^2 + \lambda_1 \|\mathbf{w}\|_2^2 + \lambda_2 \|\mathbf{w}\|_1 \right) \text{ subject to } \mathbf{w} \geq 0 \,, \quad (1)$$

where $\lambda_1 > 0$, $\lambda_2 > 0$ are respectively the ℓ_1-norm and ℓ_2-norm regularization parameters, and β denotes the intercept term. The non-negativity constraint for $\mathbf{w}$ may result in a worse performing model for the source country, but, at the same time, makes the weight transfer from a source to a target country more comprehensible (positive weights are easier to interpret) and eventually more accurate in terms of performance (see Section 4.2).

Due to the seasonal nature of influenza, our dataset of candidate queries contains a significant number of confounders, i.e. queries with frequencies that are correlated to ILI rates, but have no link to flu, such as '*college basketball*' or '*spring break*'. To remove these unrelated queries we applied a semantic filter based on word embedding representations, similar to the one proposed in [38, 72, 73]. Word embeddings were trained on the English Wikipedia corpus using the `fastText` method [12]. A topic about flu, $\mathcal{T}$, was defined as a simple set of two flu-related terms, $\mathcal{T} = \{$'flu', 'fever'$\}$. For each of the source queries, we calculate a similarity score defined as the product of the cosine similarities between the embeddings of the terms in $\mathcal{T}$ and $\mathbf{e}_\mathrm{q}$, i.e.

$$g\left(\mathbf{q}, \mathcal{T}\right) = \cos\left(\mathbf{e}_\mathrm{q}, \mathbf{e}_{\mathcal{T}_1}\right) \times \cos\left(\mathbf{e}_\mathrm{q}, \mathbf{e}_{\mathcal{T}_2}\right) \,, \quad (2)$$

where each cosine similarity component is mapped to $[0, 1]$ via $\left(\cos(\cdot, \cdot) + 1\right)/2$.[5] Queries from the source domain with $g \leq .5$ are

[5] This resolves misleading similarity scores based on different sign combinations.

Table 1: Mean ratio of query frequency over ILI rate (and standard deviation of the mean) in four countries.

Search queries	US	FR	ES	AU
flu (US/AU), grippe (FR), gripe (ES)	.036 (.010)	.033 (.012)	.032 (.011)	.031 (.016)
symptoms of flu (US/AU), symptômes de la grippe (FR), síntomas de gripe (ES)	.030 (.009)	.031 (.012)	.029 (.009)	.027 (.014)
flu in children (US/AU), grippe chez le bébé (FR), gripe en el bebé (ES)	.017 (.007)	.020 (.008)	.019 (.009)	.022 (.010)

filtered out and are not considered in our experiments. The remaining queries are used to train an elastic net. This operation further reduces the selected queries to a subset Q_S, i.e. the ones that have been allocated a nonzero weight.

3.2.2 Step 2 — Mapping source to target queries. The identified and weighted set of search queries in the source domain (Q_S) should be mapped to a set of queries in the target domain from a potential pool of target query candidates ($\mathcal{P}_T$). Queries about the same topic may vary in their textual formulation, especially when they are issued by users located in different countries. Even in cases, where countries share the same language, cultural and socioeconomic differences may result into different querying preferences. Thus, simple approaches, where search queries from the source country are translated or directly mapped to queries in the target country, are not effective.[6] In our approach, we utilize word embeddings (mono- or cross-lingual) to map source to target queries based on their broad semantic relationship. We consider both one-to-one and one-to-many query mappings from the source to the target domain. In addition, the weight associated with each source query reflects on how correlated the query is with the modeled disease rate. Therefore, another desired property is to map source queries to target ones based on their pairwise temporal correlation as this may enhance the statistical relevance of the mapping. Consequently, there is a trade-off between mapping based on semantic similarity and based on the similarity in temporal correlation. To capture both, we define a combined similarity metric, Θ, that is the weighted sum of a semantic similarity Θ_s and a correlation similarity, Θ_c, i.e.

$$\Theta = \gamma\Theta_s + (1-\gamma)\Theta_c\,, \tag{3}$$

where $\gamma \in [0, 1]$ controls the relative weighting of each. When $\gamma = 1$ the mapping is based only on semantic similarity. Conversely, when $\gamma = 0$ the mapping is based only on the correlation similarity.

Semantic similarity (Θ_s)**.** If the source and target domains have different languages, a translation module is required. For this purpose, we deploy cross-lingual word embeddings. Cross-lingual embeddings are trained using corpora from multiple languages, and can be used to compute word similarities in different languages [57, 60, 61]. Empirical evidence indicates that they can also facilitate better knowledge transfer between languages [2, 44, 47]. The majority of cross-lingual word embedding models are trained by exploiting sources of monolingual text alongside a smaller cross-lingual corpus of aligned text [56]. The alignment can be made at word [2, 5, 18, 41, 57, 60], sentence [39, 75], and document level [44, 62]. In this paper, we utilize a method for learning bilingual word embeddings proposed by Smith *et al.* [57].

First, for each of the source and target languages, we respectively learn a word embedding space based on monolingual text. For all languages considered in our experiments (English, French and Spanish) we obtained word embeddings by applying `fastText` on corresponding Wikipedia corpora [12].[7] The dimensionality of the word embeddings was set to $d = 300$. Then, we used a core selection of exact translation pairs ($\sigma \rightarrow \tau$) from the source to the target domain language to generate bilingual embeddings. Given the embedding matrices of this alignment dictionary, $\mathbf{E}_\sigma$ and $\mathbf{E}_\tau$ both $\in \mathbb{R}^{m\times d}$, where m, d denote the number of translation pairs and the dimensionality of the word embedding respectively, we learn a transformation matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ such that $\mathbf{E}_\tau \approx \mathbf{E}_\sigma\mathbf{W}$. $\mathbf{W}$ is an orthogonal matrix learned by minimizing the squared Euclidean distance between $\mathbf{E}_\sigma$ and $\mathbf{E}_\tau$, i.e.

$$\underset{\mathbf{W}}{\operatorname{argmin}}\ \|\mathbf{E}_\sigma\mathbf{W} - \mathbf{E}_\tau\|_2^2\,,\ \text{subject to } \mathbf{W}^\top\mathbf{W} = \mathbf{I}\,. \tag{4}$$

The orthogonality constraint ensures that the transformation works both ways, that is $\mathbf{E}_\tau \approx \mathbf{E}_\sigma\mathbf{W}$, $\mathbf{E}_\sigma \approx \mathbf{E}_\tau\mathbf{W}^\top$, and $\mathbf{E}_\tau \approx \mathbf{E}_\tau\mathbf{W}^\top\mathbf{W}$ [57]. In addition, Artexte *et al.* have empirically shown that it also improves the performance of machine translation [4]. The exact solution of Eq. 4 is given by $\mathbf{W} = \mathbf{V}\mathbf{U}^\top$, where $\mathbf{E}_\tau^\top\mathbf{E}_\sigma = \mathbf{U}\Sigma\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{E}_\tau^\top\mathbf{E}_\sigma$ [4, 23].

A query's embedding is defined as the average of the embeddings of its tokens, an effective practice for short texts [8, 42, 66, 72]. We denote with $\mathbf{v}_{S_i}, \mathbf{v}_{T_j}$ both $\in \mathbb{R}^{1\times d}$, the embeddings of a source query (from Q_S) and of a target query from $\mathcal{P}_T$, respectively. Then, an element ω_{ij} from the cosine similarity matrix $\Omega \in \mathbb{R}^{s\times|\mathcal{P}_T|}$ between the embeddings of source and valid target queries is given by $\omega_{ij} = \left(\mathbf{v}_{S_i}\mathbf{W}\,\mathbf{v}_{T_j}^\top\right) / \left(\|\mathbf{v}_{S_i}\mathbf{W}\|_2\|\mathbf{v}_{T_j}\|_2\right)$. Note that the cosine similarities are computed after projecting the embeddings of the source domain to the target domain using the transformation matrix $\mathbf{W}$.

In theory, we can directly use ω_{ij} to determine the k most similar target queries to the source query, thus providing a one-to-many mapping. However, in practice when conducting translations based on cross-lingual word embeddings, this may result in the presence of "hubs", i.e. target words or queries that are similar to unrealistically many different source words, a development that reduces the performance of translation [18, 57]. Smith *et al.* mitigate this effect by using an inverted softmax ranking, described next [57].

Given q_i in the source language, its translation is determined by finding candidate target queries q'_j that maximize the probability defined by

$$P_{j\rightarrow i} = \frac{\exp\left(\eta\,\omega_{ij}\right)}{\alpha_j \sum_{z=1}^{s} \exp\left(\eta\,\omega_{iz}\right)}\,, \tag{5}$$

where α_j is a normalization factor that ensures $P_{j\rightarrow i}$ is a probability, and s is the number of source queries in the vocabulary. The inverted softmax estimates the probability $P_{j\rightarrow i}$ that a candidate target query translates back to the source query, rather than

[6]We have empirical evidence about this, obtained during the first stages of this work.

[7]The embeddings were obtained from `github.com/facebookresearch/fastText`

the other way around, $P_{i \to j}$ [18, 57]. If a target query is a hub, then the denominator in Eq. 5 will be large, preventing this target query from being selected. The parameter η is learned by maximizing the log probability over the alignment dictionary ($\sigma \to \tau$), i.e., $\text{argmax}_\eta \sum_{\text{pairs } ij} \ln \left(P_{j \to i}\right)$. The top-$k$ queries from $\mathcal{P}_\text{T}$ with the highest pairing probability ($P_{j \to i}$) are then selected as possible translations of the source query q_i. Finally, we compute the semantic (cosine) similarity score Θ_s between the source query q_i and the target query q_j using $\Theta_\text{s}(q_i, q_j) = \left(\mathbf{e}_{q_i} \mathbf{W} \, \mathbf{e}_{q_j}^\top\right) / \left(\|\mathbf{e}_{q_i} \mathbf{W}\|_2 \|\mathbf{e}_{q_j}\|_2\right)$, where $\mathbf{e}_{q_i}, \mathbf{e}_{q_j}$ are the embeddings of q_i, q_j, respectively. Our experiments report results for a variety of values of k.

If the language in the source and the target domain is the same, the previously described approach is not applicable. Given potential differences in querying preferences across different countries, some of the source queries, $\mathcal{Q}_\text{S}$, may not be present in the pool of candidate target queries, $\mathcal{P}_\text{T}$. Therefore, we use cosine similarity to map each source query to the k most similar target ones using the common word embedding space for the shared language.

Temporal correlation similarity (Θ_c)**.** We compute the Pearson correlation between the frequency time series of the source and target queries over a fixed period (set to 5 years in our experiments). Since the flu season may be offset in the target domain with respect to the source domain, we computed the maximum correlation between these two frequency time series using a shifting window of $\pm\xi$ weeks. The range of possible values for ξ is determined based on the seasonal offset between the source and target countries (see Section 4). Given a source query, q_i, and a target query, q_j which is a member of a mapping set $\mathcal{T}_i$ (consisting of $k \geq 1$ queries from $\mathcal{P}_\text{T}$), and their associated daily search frequencies, $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$, respectively, the temporal correlation similarity, Θ_c, is given by

$$\Theta_\text{c}(q_i, q_j) = \rho\left(\mathbf{x}_i(t), \mathbf{x}_j(t + l_{ij})\right) , \quad (6)$$

where $\rho(x_i(t), x_j(t + l_{ij}))$ denotes the optimal Pearson correlation coefficient between $\mathbf{x}_i, \mathbf{x}_j$ within the shifting window. Note that the optimal window is independently computed for each target query in $\mathcal{T}_i$, and thus optimal shifts may vary.

3.2.3 Step 3 — Weighting target queries. In the previous steps, we have established that a source query q_i, which has received a regression weight w_i, is mapped to a set, $\mathcal{T}_i$, of $k \geq 1$ queries in the target domain. If $k = 1$, then we can directly assign w_i to the single target query. If $k > 1$, then the source query's weight, w_i, should be distributed across these k mapping target queries. To perform this, we have considered two alternatives:

- **Uniform.** We divide the source query weight, w_i, by the number of queries q'_j in $\mathcal{T}_i$, and assign each query in $\mathcal{T}_i$ a weight equal to $w'_j = w_i/k$.
- **Non-uniform.** The k target query weights are determined based on each target query's similarity score Θ_{ij}, $j \in \{2, \ldots, k\}$, with the source query (see Eq. 3). More specifically, a target weight w'_j is defined as $w'_j = w_i \Theta_{ij'} / \sum_{q'_j \in \mathcal{T}_i} \Theta_{ij'}$.

To obtain a baseline performance estimate, we randomly shuffle the established query mappings in Step 2, and then transfer the source weights to k target queries using the uniform approach. We repeat this process multiple times and report the mean performance of these randomized transfer learning models.

4 EXPERIMENTS

We deploy the proposed transfer learning framework to estimate ILI rates in three target countries without using any ground truth from these countries to supervise modeling. US is always set as the source country, while the target countries are FR, ES and AU. We assess the performance of the proposed model, comparing it to various baselines, and also provide a qualitative analysis, aiming to interpret some of the intrinsic properties of our approach.

Settings. After applying the semantic filter (Eq. 2) to the pool of 34,121 US queries, 1,403 queries were retained. The applied evaluation protocol is as follows. We train a source model (US) using the first 5 flu seasons (2007-12). A flu season is conventionally defined as the 1-year long period from the first week in September to the last week of August in the next year.[8] Prior to applying elastic net, we maintain search queries that have a $\geq$ .3 Pearson correlation with the US ILI rates (these queries may vary per training fold). We then transfer the model to FR, ES, and AU and test it in the following flu season (2012-13). Then, we move our training data window to include the 2012-13 flu season and remove the first flu season (2007-08), and test in the following season (2013-14), so that we still have 5 flu seasons to train. We repeat this process until we have tested on the last flu season in our data set (2015-16), evaluating performance 4 times in total. The window size (ξ) used for identifying optimal correlations between the frequency time series of the source and target queries (see Section 3) is set to $\pm$6 weeks for FR and ES. The window is the same for AU, although prior to applying it, the query frequency time series are shifted by 6 months to account for the seasonal difference in the northern and southern hemispheres. For the one-to-k mapping from a source to a set of target queries, we explore sizes up to $k = 5$ (values > 5 did not yield any different insights). We measure the performance of transferred models by comparing our estimates with their national public health estimates, using Pearson correlation (r), mean absolute error (MAE), and root mean squared error (RMSE). Regression errors are computed after reverting inferences back to their corresponding non standardized values.

Baseline models. To demonstrate the effectiveness of our transfer learning framework, we compare it with four baseline models:

- **Random**. After determining the mapping between source and target queries, the pairs (one-to-k) are randomly permuted. The source query weight is uniformly distributed across the mapped k target queries. We repeat this process 2,000 times and report the average inference performance. This random assignment of query weights provides a possibly worst case baseline.
- **Transfer component analysis (TCA)**. TCA is a transfer learning approach that aims to learn transfer components across source and target domains in a reproducing kernel Hilbert space using maximum mean discrepancy [48]. After we map source to target queries, TCA is applied to source and target query frequencies.
- **Unsupervised query selection based on semantic similarity**. We apply a semantic filter (described in Eq. 2) to remove queries that are irrelevant to the flu topic. The term pairs {'grippe',

[8]Note that for AU this may result into including the end of a flu season and the beginning of the next in training and test folds.

Table 2: Performance estimates for the US→FR transfer learning task. Different values of γ determine how queries are mapped from the source to the target domain (γ=1: semantic similarity only, γ=0: temporal correlation only, $\gamma \in (0, 1)$: joint similarity score). Numbers in parentheses represent the standard deviation of the error. The best performance among all transfer learning models is denoted in bold. The best performance among models under a common γ value is underlined. Only the best random mapping performance (R) is enumerated per choice of γ. The last two rows show the performance of the baseline models.

Mapping	k	w	09/2012 – 09/2013			09/2013 – 09/2014			09/2014 – 09/2015			09/2015 – 09/2016			Average		
			r	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**
$\gamma = 0$	1	–	.797	78.905	136.098	.789	59.584	93.752	.900	56.107	92.324	.855	51.533	78.073	.835 (.045)	61.532 (10.429)	100.062 (21.690)
	2	U	.803	80.044	137.247	.794	59.961	94.853	.890	58.372	96.282	.843	55.532	84.438	.833 (.038)	63.477 (9.696)	103.205 (20.179)
	3	U	.802	79.010	135.905	.796	59.750	95.350	.896	57.241	94.451	.844	57.306	86.588	.834 (.040)	63.327 (9.111)	103.073 (19.260)
	4	U	.798	79.077	135.892	.795	59.529	95.295	.895	58.380	95.852	.834	59.729	90.180	.830 (.040)	64.179 (8.617)	104.305 (18.370)
	5	U	.799	78.881	135.743	.794	58.508	95.036	.893	58.439	96.988	.829	60.075	91.182	.829 (.040)	63.976 (8.630)	104.737 (18.023)
	2	NU	.803	80.012	137.180	.794	59.971	94.869	.891	58.360	96.268	.843	55.502	84.399	.833 (.038)	63.461 (9.689)	103.179 (20.159)
	3	NU	.802	78.999	135.881	.796	59.763	95.360	.896	57.244	94.453	.844	57.271	86.538	.834 (.040)	63.319 (9.110)	103.058 (19.259)
	4	NU	.799	79.068	135.875	.795	59.519	95.278	.895	58.367	95.834	.834	59.676	90.106	.830 (.040)	64.157 (8.623)	104.273 (18.381)
	5	NU	.799	78.868	135.725	.794	58.499	95.015	.893	58.434	96.972	.829	60.029	91.110	.829 (.040)	63.957 (8.632)	104.706 (18.033)
	1	R	.771	125.422	152.275	.731	93.122	105.769	.807	138.579	158.000	.825	102.972	113.607	.783 (.036)	115.024 (17.943)	132.413 (22.982)
$\gamma = 1$	1	–	.964	51.885	77.728	.928	24.373	35.801	.974	51.623	69.254	.917	75.416	92.946	.946 (.024)	50.824 (18.071)	68.932 (20.927)
	2	U	.967	41.298	68.164	.939	22.993	33.287	.973	62.869	81.119	.924	84.469	102.422	.951 (.020)	52.907 (23.049)	71.248 (25.099)
	3	U	.967	39.789	67.336	.947	21.219	30.446	.972	58.654	79.471	.933	76.235	93.338	.955 (.016)	48.974 (20.564)	67.648 (23.366)
	4	U	.965	40.120	65.882	.947	24.037	33.095	.970	63.290	85.390	.939	77.601	93.301	.955 (.013)	51.262 (20.638)	69.417 (23.224)
	5	U	.965	37.632	61.217	.952	26.136	35.651	.972	66.825	90.248	.943	78.479	93.855	.958 (.011)	52.268 (21.190)	70.243 (23.642)
	2	NU	.968	41.272	68.016	.939	22.925	33.213	.973	61.971	80.280	.924	83.058	101.160	.951 (.020)	52.306 (22.495)	70.667 (24.658)
	3	NU	.967	39.665	66.933	.948	21.189	30.378	.973	58.568	79.476	.933	75.661	92.917	.955 (.016)	48.770 (20.388)	67.426 (23.280)
	4	NU	.966	39.754	65.480	.948	23.794	32.767	.971	62.957	85.275	.939	76.868	92.866	.956 (.013)	50.843 (20.486)	69.097 (23.236)
	5	NU	.966	37.295	60.749	.952	25.925	35.383	.972	66.890	90.583	.943	77.969	93.647	.958 (.012)	52.020 (21.167)	70.091 (23.805)
	3	R	.891	83.535	113.537	.890	79.396	86.904	.949	116.532	124.478	.922	109.746	119.219	.913 (.024)	97.302 (16.084)	111.034 (14.459)
	2	C	.968	39.972	65.695	.941	21.639	31.190	.974	59.103	77.964	.926	78.798	97.444	.952 (.019)	49.878 (21.313)	68.073 (24.117)
	3	C	.967	38.062	64.349	.949	20.408	29.002	.973	56.188	77.822	.933	72.492	90.289	.956 (.016)	46.788 (19.501)	65.365 (22.911)
	4	C	.965	38.225	63.063	.949	22.869	31.161	.971	60.623	83.764	.938	73.644	90.367	.956 (.013)	48.840 (19.629)	67.089 (23.059)
	5	C	.966	35.827	58.820	.953	24.940	33.619	.973	63.562	87.764	.942	74.547	90.793	.958 (.012)	49.719 (20.094)	67.749 (23.325)
$\gamma_{\mathrm{opt}} = .5$	1	–	.968	33.475	53.775	.951	22.615	34.416	.973	34.793	58.007	.944	45.324	62.417	**.959 (.012)**	**34.052 (8.043)**	**52.153 (10.687)**
	2	U	.959	37.461	60.529	.939	24.885	38.056	.967	43.197	69.883	.930	54.504	74.766	.949 (.015)	40.012 (10.671)	60.809 (14.097)
	3	U	.954	38.786	63.909	.939	26.390	39.771	.968	44.241	71.312	.931	61.182	81.592	.948 (.014)	42.650 (12.503)	64.146 (15.410)
	4	U	.948	41.150	69.125	.934	29.553	43.996	.966	47.021	74.662	.932	62.330	82.811	.945 (.014)	45.014 (11.810)	67.649 (14.498)
	5	U	.945	41.936	71.322	.925	30.387	46.164	.963	46.108	75.703	.931	61.750	82.670	.941 (.015)	45.045 (11.233)	68.965 (13.772)
	2	NU	.959	37.414	60.456	.939	24.881	38.036	.967	43.118	69.763	.930	54.329	74.599	.949 (.015)	39.936 (10.610)	60.714 (14.045)
	3	NU	.954	38.675	63.792	.940	26.423	39.789	.968	44.452	71.495	.931	61.147	81.601	.948 (.014)	42.674 (12.495)	64.169 (15.428)
	4	NU	.948	40.867	68.727	.935	29.381	43.748	.966	47.093	74.691	.932	62.323	82.804	.945 (.014)	44.916 (11.890)	67.492 (14.591)
	5	NU	.946	41.610	70.892	.926	30.201	45.863	.963	46.192	75.685	.931	61.788	82.685	.942 (.015)	44.948 (11.333)	68.781 (13.881)
	1	R	.913	86.752	110.096	.846	72.130	83.158	.943	94.681	109.176	.942	97.352	104.952	.911 (.039)	87.729 (9.813)	101.845 (10.962)
Unsupervised	–	–	.936	–	–	.870	–	–	.947	–	–	.910	–	–	.916 (.030)	–	–
Supervised	–	–	.977	27.331	50.643	.979	23.665	33.994	.992	34.345	62.803	.987	15.011	21.956	.984 (.006)	25.088 (6.970)	42.349 (15.595)

k: number of target queries (1-to-k mapping), **w**: weighting approach, **U**: uniform, **NU**: non-uniform, **C**: correlation, **R**: random

'fièvre'}, {'gripe', 'fiebre'} and {'flu', 'fever'} are used to define this semantic filter in FR, ES and AU, respectively. Queries with $g \leq .5$ are filtered out and are not considered in our experiments. The mean weekly frequency of the retained queries is regarded as a proxy of the estimated ILI rates. These estimates are in different scale with the true ILI rates, thus we only report their Pearson correlation (r).

- **Supervised learning**. We first apply a semantic filter (see point above) to the queries of each target country. We then train an elastic net, after maintaining only queries that have a moderate correlation with the ground truth ($r \geq .3$ with the target values in the training data). This is inline with previously proposed, state-of-the-art supervised models for the task [38] and is considered as the top performance we could obtain, if we had access to ground truth in the target countries.

4.1 Quantitative analysis

Performance estimates are enumerated in Tables 2, 3, and 4 for each transfer learning task (US→FR, US→ES, US→AU). We first explored the extreme cases of $\gamma = 0$ and $\gamma = 1$ (Eq. 3) that result to using only temporal correlation or semantic similarity, respectively.

For $\gamma = 0$, spurious queries could be included in the target domain's mappings. This is a result of the way the pool of target queries, $\mathcal{P}_{\mathrm{T}}$, was originally formed (see Section 2). Seasonal search queries, correlating with the occurrence of flu incidents in a population, are very likely to be selected as mappings, e.g. "symptoms flu" was mapped to "ski serre chevalier" in the US→FR task. Seasonal activities or expressions may change in time, and thus such queries are very unstable predictors. In fact, the best average performance we can obtain for $\gamma = 0$ is considerably worse (MAEs of 61.532, 25.977 and 42.348 for FR, ES, and AU) than for alternative values. Setting $k = 1$ provides the best results on average. In general, performance is not affected much by different choices of weighting (uniform, non-uniform) or the number of queries in a mapping (k).

For $\gamma = 1$, we obtain on average more accurate estimates than for $\gamma = 0$. As a precursor to the joint similarity, we also introduce a correlation-based weighting scheme (denoted by "C"), which uses the optimal correlation between source and target queries (after

Table 3: Performance estimates for US→ES transfer learning task. Please refer to Table 2's caption for further information.

Mapping	k	w	09/2012 – 09/2013			09/2013 – 09/2014			09/2014 – 09/2015			09/2015 – 09/2016			Average		
			r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE
$\gamma = 0$	1	–	.808	25.068	41.104	.807	25.789	42.137	.843	29.221	47.360	.791	25.134	38.497	.812 (.019)	26.303 (1.708)	42.275 (3.222)
	2	U	.799	25.589	42.631	.843	23.850	39.092	.844	30.069	48.120	.821	24.470	36.902	.827 (.019)	25.994 (2.434)	41.686 (4.240)
	3	U	.795	25.756	42.883	.840	23.669	38.934	.843	29.509	48.189	.813	24.989	37.713	.823 (.020)	25.981 (2.169)	41.930 (4.088)
	4	U	.783	26.504	43.662	.835	23.671	39.207	.844	29.850	48.335	.809	25.715	38.745	.818 (.024)	26.435 (2.226)	42.487 (3.884)
	5	U	.783	26.579	43.391	.840	23.605	38.861	.842	30.336	48.800	.806	26.586	39.843	.818 (.025)	26.776 (2.388)	42.724 (3.892)
	2	NU	.799	25.579	42.610	.843	23.852	39.095	.844	30.060	48.111	.821	24.472	36.907	.827 (.019)	25.991 (2.430)	41.681 (4.233)
	3	NU	.795	25.748	42.867	.840	23.670	38.936	.843	29.503	48.176	.813	24.989	37.712	.823 (.020)	25.977 (2.167)	41.922 (4.082)
	4	NU	.784	26.491	43.643	.835	23.671	39.209	.844	29.842	48.325	.809	25.708	38.734	.818 (.023)	26.428 (2.223)	42.478 (3.881)
	5	NU	.783	26.567	43.380	.840	23.605	38.866	.842	30.324	48.785	.806	26.575	39.826	.818 (.025)	26.768 (2.384)	42.714 (3.887)
	3	R	.830	40.548	46.584	.903	36.241	40.718	.846	53.929	61.098	.813	45.762	49.637	.848 (.034)	44.120 (6.591)	49.509 (7.419)
$\gamma = 1$	1	–	.954	28.614	34.944	.976	27.777	30.129	.919	44.638	50.082	.899	43.761	46.590	.937 (.030)	36.197 (8.013)	40.436 (8.175)
	2	U	.955	27.342	33.979	.976	27.118	29.294	.923	44.723	50.213	.925	44.518	49.547	**.945 (.022)**	35.926 (8.696)	40.758 (9.274)
	3	U	.958	25.523	31.885	.971	28.293	32.055	.916	47.603	53.909	.917	48.513	54.053	.941 (.024)	37.483 (10.625)	42.975 (11.006)
	4	U	.960	25.316	31.623	.973	27.998	31.797	.918	46.862	53.458	.918	47.443	52.823	.942 (.025)	36.905 (10.294)	42.425 (10.718)
	5	U	.957	24.445	30.821	.975	27.169	30.959	.917	45.775	52.620	.914	45.854	51.505	.941 (.026)	35.811 (10.050)	41.476 (10.594)
	2	NU	.955	26.336	32.978	.977	26.069	28.232	.923	43.543	49.056	.925	43.389	48.409	**.945 (.022)**	34.834 (8.632)	39.669 (9.221)
	3	NU	.958	25.532	31.879	.971	28.327	32.076	.917	47.471	53.737	.917	48.356	53.908	.941 (.024)	37.422 (10.543)	42.900 (10.923)
	4	NU	.960	25.324	31.587	.973	28.020	31.814	.919	46.770	53.334	.917	47.391	52.769	.942 (.025)	36.876 (10.251)	42.376 (10.678)
	5	NU	.958	24.432	30.759	.975	27.197	30.990	.917	45.778	52.576	.915	45.951	51.574	.941 (.026)	35.839 (10.073)	41.475 (10.607)
	2	R	.731	47.277	53.345	.804	44.924	52.394	.795	60.370	70.934	.719	48.986	56.506	.762 (.038)	50.389 (5.940)	58.295 (7.454)
	2	C	.954	25.520	33.516	.976	24.408	26.693	.923	41.827	47.782	.924	41.142	46.278	.944 (.022)	33.224 (8.273)	38.567 (8.816)
	3	C	.957	23.642	31.398	.970	25.353	29.090	.916	44.358	50.846	.916	45.405	51.174	.940 (.024)	34.690 (10.217)	40.627 (10.416)
	4	C	.960	23.339	30.912	.973	24.900	28.709	.919	43.431	50.236	.918	44.297	49.873	.942 (.025)	33.992 (9.892)	39.933 (10.153)
	5	C	.957	24.137	30.513	.974	26.555	30.466	.917	44.598	51.464	.915	45.662	51.359	.941 (.026)	35.238 (9.936)	40.950 (10.461)
$\gamma_{\text{opt}} = .2$	1	–	.931	21.419	30.004	.948	15.403	23.900	.907	27.050	39.864	.888	26.762	35.420	.918 (.023)	**22.658 (4.751)**	**32.297 (5.974)**
	2	U	.926	21.433	29.944	.941	17.334	25.525	.899	30.166	43.243	.877	30.662	40.272	.911 (.025)	24.899 (5.705)	34.746 (7.260)
	3	U	.936	21.249	28.841	.961	18.189	24.028	.908	31.608	42.568	.900	35.661	43.995	.926 (.024)	26.677 (7.186)	34.858 (8.608)
	4	U	.945	21.016	28.161	.965	18.720	23.647	.917	32.235	41.483	.910	37.141	44.448	.934 (.022)	27.278 (7.654)	34.435 (8.742)
	5	U	.946	20.977	28.041	.967	18.727	23.321	.910	33.330	43.018	.903	36.846	44.547	.932 (.026)	27.470 (7.760)	34.732 (9.219)
	2	NU	.926	21.427	29.932	.941	17.321	25.510	.899	30.135	43.214	.877	30.626	40.233	.911 (.025)	24.877 (5.694)	34.723 (7.251)
	3	NU	.936	21.254	28.845	.961	18.186	24.037	.908	31.583	42.554	.900	35.629	43.969	.926 (.024)	26.663 (7.171)	34.851 (8.595)
	4	NU	.945	21.023	28.158	.965	18.739	23.682	.917	32.241	41.509	.910	37.128	44.438	.934 (.022)	27.283 (7.643)	34.447 (8.734)
	5	NU	.946	20.983	28.033	.967	18.747	23.364	.910	33.337	43.028	.903	36.872	44.568	.932 (.026)	27.485 (7.762)	34.748 (9.215)
	1	R	.865	32.859	40.942	.931	33.323	38.687	.878	49.262	57.762	.814	45.799	51.424	.872 (.042)	40.311 (7.325)	47.204 (7.763)
$\gamma = .5$	1	–	.945	20.016	28.161	.965	17.720	23.647	.917	31.235	41.483	.910	36.141	44.448	.934 (.022)	26.278 (7.654)	34.435 (8.742)
Unsupervised	–	–	.936	–	–	.976	–	–	.910	–	–	.878	–	–	.925 (.036)	–	–
Supervised	–	–	.968	19.788	24.487	.993	30.642	41.059	.972	24.779	35.861	.954	13.271	20.992	.971 (.014)	22.120 (6.392)	30.600 (8.166)

k: number of target queries (1-to-k mapping), **w**: weighting approach, **U**: uniform, **NU**: non-uniform, **C**: correlation, **R**: random

deploying a shifting window) to determine the proportion of the source weight that will be allocated to the k mapped queries. In countries that deploy a translation module based on bilingual word embeddings, the "C" scheme ($k = 2$ or 3) outperforms the other two (uniform, non-uniform). For the US→AU task, where high semantic similarity often means that very similar queries are being mapped to each other (given the common language), the optimal model is obtained for $k = 1$, and thus, no further distribution of the weights is required. With or without the "C" weighting scheme, better performance is achieved compared to setting $\gamma = 0$ (MAEs of 46.788/48.77, 33.224/34.834 and 34.509/30.275 for FR, ES, and AU).

The joint similarity scheme attempts to combine the positive attributes of semantic and correlation based similarities. To assess its potential contribution, we performed a grid search using 9 values of γ (from .1 to .9), and presented the results for the best performing one (γ_{opt}). For completeness, we also show results for the default choices of $\gamma = .5$ and $k = 1$. Firstly, the application of the joint similarity leads to significant performance improvements in all tasks (MAEs of 34.052, 22.658 and 22.043 for FR, ES, and AU). Secondly, the best performing model consistently occurs for $k = 1$, i.e. for one-to-one query mappings, where no weight redistribution is required. Finally, although results do not deviate much from the default settings of $\gamma = .5$ and $k = 1$, there are discrepancies between the optimal γ value for each task ($\gamma_{\text{opt}} = .5$, .2 and .9 for FR, ES, and AU). One possible explanation may be that this is an artefact of the intrinsic characteristics (size, semantic/temporal similarities) of the pool of candidate target queries used for each task (see Section 4.2).

Better performance is always obtained (in terms of MAE and RMSE) compared to the random mapping allocation baseline ("R"), the best performance estimates of which per γ value are provided. The same holds for TCA, which performs even worse than random (results are omitted). One explanation for this is that TCA fails to capture the time series structure of this particular data set, an essential property for producing a meaningful solution. Furthermore, the optimal models (joint similarity) outperform the unsupervised baseline in terms of correlation, the only metric which is relevant in this case. Finally, compared to the fully supervised elastic net, the transfer learning unsupervised approach reaches to a comparable performance, which is worse by 23.15%, 5.55%, and 17.5% (in terms of RMSE), for FR, ES, and AU, respectively.

Fig. 2 plots the time series of a selection of these estimates, including the ones of the best performing models, in comparison to the ground truth, for each target country. We can see how estimates become significantly better when the joint similarity is

Table 4: Performance estimates for the US→AU transfer learning task. Please refer to Table 2's caption for further information.

Mapping	k	w	09/2012 – 09/2013			09/2013 – 09/2014			09/2014 – 09/2015			09/2015 – 09/2016			Average		
			r	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**
$\gamma = 0$	1	—	.704	38.804	50.140	.677	39.151	48.508	.630	51.412	65.215	.787	40.025	57.421	.700 (.056)	42.348 (5.359)	55.321 (6.830)
	2	U	.622	41.824	55.943	.663	41.708	50.752	.633	52.017	66.448	.763	40.557	59.312	.670 (.055)	44.027 (4.734)	58.114 (5.873)
	3	U	.621	42.263	56.819	.669	42.900	51.487	.631	53.041	67.754	.769	41.330	59.468	.672 (.058)	44.883 (4.840)	58.882 (6.055)
	4	U	.607	42.040	56.755	.669	42.501	51.008	.634	51.868	66.404	.759	40.287	58.660	.667 (.056)	44.174 (4.611)	58.207 (5.678)
	5	U	.600	41.900	56.618	.671	41.950	49.692	.647	50.958	64.744	.761	40.899	58.979	.670 (.058)	43.927 (4.164)	57.508 (5.561)
	2	NU	.623	41.886	55.947	.663	41.642	50.818	.633	52.068	66.590	.763	40.617	59.384	.670 (.055)	44.053 (4.747)	58.185 (5.908)
	3	NU	.620	42.263	56.812	.668	42.857	51.533	.631	53.062	67.852	.769	41.373	59.540	.672 (.058)	44.889 (4.845)	58.934 (6.081)
	4	NU	.607	42.031	56.745	.669	42.466	51.039	.634	51.909	66.504	.759	40.343	58.732	.667 (.056)	44.187 (4.621)	58.255 (5.708)
	5	NU	.600	41.885	56.601	.671	41.928	49.723	.647	51.011	64.844	.761	40.935	59.032	.670 (.058)	43.940 (4.186)	57.550 (5.589)
	1	R	.653	60.835	71.392	.710	52.090	62.045	.628	67.895	78.856	.738	69.695	75.320	.683 (.043)	62.629 (7.069)	71.903 (6.468)
$\gamma = 1$	1	—	.916	23.447	26.436	.871	13.994	18.129	.902	35.315	42.126	.971	48.344	50.617	.915 (.035)	30.275 (13.143)	34.327 (13.150)
	2	U	.900	28.828	33.029	.880	18.583	22.656	.925	39.274	45.149	.989	59.174	60.026	.923 (.040)	36.465 (15.320)	40.215 (14.366)
	3	U	.896	30.804	35.148	.881	19.492	23.743	.938	36.748	42.294	.990	57.829	58.516	.926 (.041)	36.218 (14.216)	39.925 (12.999)
	4	U	.889	30.876	35.549	.872	21.475	26.089	.935	37.484	42.966	.994	57.871	58.397	.922 (.047)	36.926 (13.636)	40.750 (12.180)
	5	U	.882	31.248	35.738	.868	21.320	25.883	.936	37.615	43.059	.992	58.773	59.318	.919 (.047)	37.239 (14.002)	41.000 (12.584)
	2	NU	.902	28.789	32.947	.880	18.497	22.565	.925	39.278	45.150	.989	59.007	59.861	.924 (.039)	36.393 (15.287)	40.131 (14.347)
	3	NU	.897	30.805	35.137	.882	19.510	23.775	.938	36.973	42.482	.990	57.779	58.462	.927 (.041)	36.267 (14.193)	39.964 (12.978)
	4	NU	.890	30.839	35.484	.873	21.367	25.986	.936	37.554	42.999	.994	57.825	58.354	.923 (.046)	36.896 (13.655)	40.706 (12.205)
	5	NU	.884	31.217	35.678	.870	21.261	25.830	.936	37.609	43.019	.992	58.770	59.309	.920 (.047)	37.214 (14.022)	40.959 (12.603)
	1	R	.825	58.539	60.310	.793	42.200	46.818	.890	55.940	61.462	.963	65.023	66.924	.868 (.064)	55.426 (8.491)	58.878 (7.627)
	2	C	.905	27.444	31.356	.881	17.547	21.520	.925	37.373	43.387	.989	58.318	59.229	.925 (.040)	35.171 (15.399)	38.873 (14.510)
	3	C	.900	28.802	32.701	.882	18.039	22.091	.939	34.534	40.310	.990	56.660	57.516	**.928 (.041)**	34.509 (14.381)	38.154 (13.316)
	4	C	.894	28.643	32.867	.874	19.505	23.747	.938	34.613	40.360	.994	56.309	57.011	.925 (.045)	34.768 (13.828)	38.496 (12.579)
	5	C	.888	29.149	33.118	.870	19.259	23.507	.939	34.622	40.252	.992	57.220	57.962	.922 (.047)	35.063 (14.211)	38.710 (12.993)
$\gamma_{opt} = .9$	1	—	.922	11.997	14.986	.879	15.084	18.011	.898	24.898	31.110	.985	36.191	38.271	.921 (.039)	**22.043 (9.649)**	**25.594 (9.796)**
	2	U	.892	16.642	19.922	.881	15.719	19.009	.923	23.858	30.280	.988	39.919	41.175	.921 (.041)	24.034 (9.895)	27.596 (9.282)
	3	U	.890	18.641	22.543	.876	18.391	21.453	.930	23.965	29.934	.989	41.232	42.249	.921 (.043)	25.557 (9.510)	29.045 (8.549)
	4	U	.883	19.078	23.494	.866	19.766	22.757	.928	23.691	29.686	.991	40.159	41.138	.917 (.047)	25.673 (8.721)	29.269 (7.590)
	5	U	.875	20.091	24.960	.862	18.791	21.614	.933	23.474	29.474	.991	41.433	42.483	.915 (.050)	25.947 (9.288)	29.633 (8.171)
	2	NU	.894	16.565	19.826	.882	15.679	18.961	.923	23.830	30.226	.988	39.809	41.071	.922 (.040)	23.971 (9.873)	27.521 (9.270)
	3	NU	.892	18.588	22.457	.877	18.312	21.353	.930	23.995	29.967	.989	41.230	42.245	.922 (.042)	25.531 (9.534)	29.005 (8.589)
	4	NU	.885	19.043	23.410	.867	19.639	22.621	.929	23.690	29.673	.991	40.229	41.204	.918 (.047)	25.650 (8.781)	29.227 (7.665)
	5	NU	.877	19.983	24.795	.864	18.716	21.530	.933	23.414	29.390	.991	41.416	42.462	.916 (.049)	25.882 (9.318)	29.544 (8.210)
	1	R	.844	47.859	50.120	.817	37.727	40.926	.900	54.008	59.263	.940	55.980	59.071	.875 (.047)	48.893 (7.254)	52.345 (7.791)
$\gamma = .5$	1	—	.871	18.642	23.367	.848	17.735	20.735	.873	27.140	32.733	.930	39.651	43.484	.880 (.298)	25.792 (8.982)	30.080 (9.208)
Unsupervised	—	—	.815	—	—	.810	—	—	.881	—	—	.942	—	—	.862 (.054)	—	—
Supervised	—	—	.891	19.353	25.297	.865	22.048	25.200	.939	18.658	22.473	.971	11.255	14.159	.916 (.041)	17.829 (4.001)	21.782 (4.545)

k: number of target queries (1-to-k mapping), **w**: weighting approach, **U**: uniform, **NU**: non-uniform, **C**: correlation, **R**: random

used versus its extremes. The transferred models can very often estimate the peak of the flu season accurately. This includes the time of occurrence as well as its intensity. Notably, ILI rates in the target countries differ in terms of scale compared to ones of the source, but the proposed models are capable of capturing different scales effortlessly, providing further evidence about the user search behavior similarities among different countries (Section 3.1). At the same time, most models show some inaccuracies, especially during the time periods with very moderate flu circulation (e.g. summer).

4.2 Qualitative analysis

A fair criticism for the proposed framework is that in a practical scenario the optimal values for γ and k cannot be validated. However, we have already demonstrated that the default settings of $\gamma = .5$ and $k = 1$ provide very satisfactory performance in all our case studies. Fig. 3 looks further into this, depicting performance estimates (MAE) for different values of γ. As discussed previously, optimal γ values differ per target country. Interestingly, all error trends are monotonically decreasing (as γ increases) until they reach a minimum, and then begin to monotonically increase. We argue that γ_{opt} reflects on the actual pool of candidate target queries ($\mathcal{P}_T$), although we have a small sample size to be able to empirically prove this. In our data, the average correlation over the average semantic similarity ratio between all source-target query pairs is equal to 1.143, .982 and 2.261, for the FR, ES, and AU tasks respectively. These ratios depend on characteristics of the target queries which we are not controlling for in our approach. They do correlate with the respective optimal γ values (.5, .2, and .9), an insight that can be used to make a more informed choice of γ in future applications of the proposed framework.

Table 5 lists the top-5 query mappings that were the most impactful in the ILI estimates on average during the 10 weeks with the lowest and greatest MAEs (for the optimal transfer models). Impact is determined by the percentage of an estimated ILI rate that is contributed by a query (frequency × weight / estimated ILI rate). The identified pairs during the weeks with the lowest errors are topically coherent (about flu) and in many occasions are accurate translations from the source to the target language. On the other hand, pairs responsible for the largest errors include inaccurate translations that sometimes lead to an off-topic target query selection. For example, "24 hour flu" is mapped to "grippe intestinale" (impact: 13.2%),[9] "child fever" to "sinusitis" (7.7%), and "child temperature" to "warmer" (9.8%). Nevertheless, it is encouraging

[9]"Grippe intestinale" translates to "stomach flu" (formally "viral gastroenteritis").

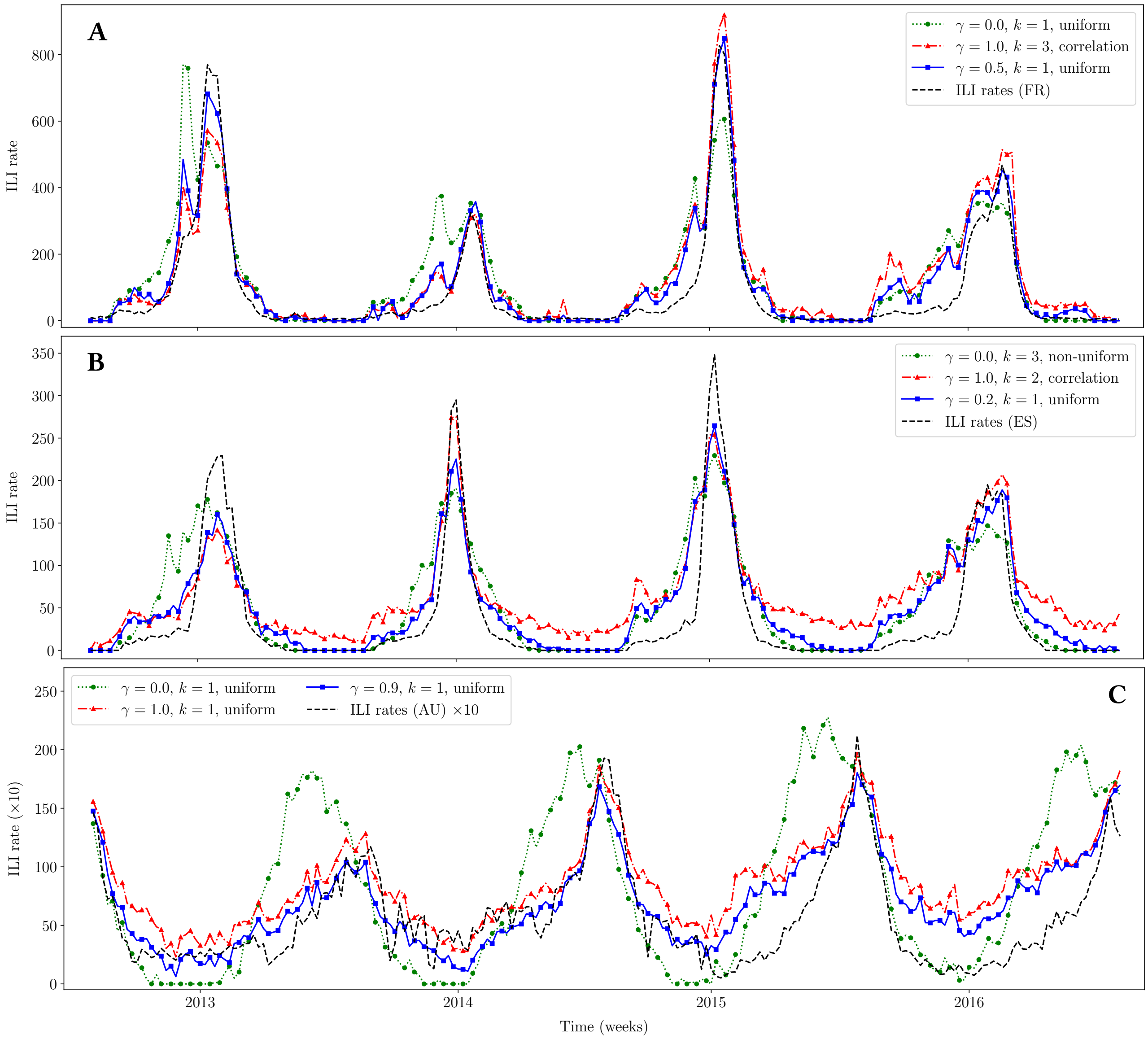

Figure 2: Comparison of transfer learning models for estimating ILI rates in France (A), Spain (B) and Australia (C) with the corresponding actual ILI rates obtained by health agencies in these countries.

that some of these mappings may have been avoided by carefully preprocessing the target query candidates to avoid spurious queries.

The optimal joint similarity transfer models do not improve by increasing the number of target queries ($k > 1$). An interpretation for that might be drawn by the fact that for $k = 1$ at most 77.9% of the selected target queries are unique (at least 22.1% are repetitive selections). Hence, the method seems to be converging to a subset of queries already for $k = 1$. As k increases, the error increases monotonically. This might be due to the existence of various spurious queries in the feature space which are being introduced as additional mappings.

Finally, the choice of adding a non-negativity constraint to the regularized regression function for the source domain (Eq. 1), was also empirically justified. When it is removed, we can learn a more accurate source model for the US, but the MAE on the target countries increases on average by 20.6%, 21.6%, and 20.5% for FR, ES, and AU respectively. This confirms our original assumption that transferring negative weights is a harder task, and thus, error-prone.

5 RELATED WORK

The fundamental properties of transfer learning have been thoroughly discussed in relevant literature [6, 7, 40, 49, 59, 65]. In contrast to traditional machine learning methods, which assume that the training and test data belong to the same domain, i.e. they are drawn from the same feature space and distribution, transfer learning aims to improve the learning function in a target domain by

Table 5: Top-5 target queries (with source mappings) in terms of mean ILI estimate impact (%) in the 10 weeks with the lowest and greatest MAE (all test periods), for all target countries (TC), based on their respective optimal transfer learning models.

TC	Mappings during accurate estimates	Mappings during inaccurate estimates
FR	flu incubation period → grippe durée (10.9), cough fever → la toux (6.3), how to treat flu → comment soigner une grippe (6), fever flu → fièvre de la grippe (5.47), flu treatment → traitement de la grippe (4.95)	24 hour flu → grippe intestinale (13.24), influenza a treatment → grippe traitement (8.07), remedies for colds → rhume de cerveau (6.75), child temperature → température du corps (6.37), child fever → fièvre adulte (6.04)
ES	symptoms of flu → symptômes grippe (9.04), fever flu → con gripe (7.49), cough fever → la tos (6.34), flu incubation period → cuanto dura una gripe (5.19), how to treat a fever → para bajar la fiebre (5.03)	mucinez for kids → tratmiento de la grippe (20.76), child fever → sinusitis (7.76), influenza a treatment → con gripe (7.02), symptoms pneumonia → bronquitis (6.04), child temperature → temperatura corporal (5.62)
AU	treatment for the flu → flu treatment (9.85), cough fever → cough and fever (8.05), flu type → influenza type (5.37), symptoms of flu → symptoms of flu (5.11), flu incubation period → flu incubation period (5.03)	24 hour flu → flu duration (11.51), child temperature → warmer (9.77), how to treat a fever → have a fever (6.94), tamiflu and breastfeeding → flu while pregnant (6.81), robitussin cf → colds (5.18)

transferring knowledge from a related, source domain. This concept has been successfully applied to various tasks, including text classification [14, 16, 22, 48], part of speech tagging [10, 28], machine translation [20, 29], and image classification [19, 30, 71].

In this work, we present a statistical framework for transferring a disease surveillance model from a source country, where supervised learning is applicable, to a target country, where no ground truth is available. We formulate it as a cross-lingual transductive regression task [49], which poses the following challenges: (a) ground truth is not available in the target domain, and (b) features (queries) may not belong in the same feature space due to linguistic or cultural differences. Due to (a), multi-task learning models, such as this solution for ILI [72], cannot be used because they still require partial ground truth from the target domain to capture the relationship between the different tasks [13]. To solve (b), a few studies have attempted to learn a mapping of both source and target languages to the same space [27, 55, 57, 64]. For example, Prettenhofer and Stein used unlabeled documents along with a word translation oracle to automatically induce task-specific, cross-lingual correspondences for cross-lingual text classification [55]. In this paper, we used cross-lingual word embeddings to align different languages [57].

Methods have also been proposed for reducing the distance between the source and target features [48, 70]. For example, Pan *et al.* proposed TCA to learn transfer components across source and target domains in a reproducing kernel Hilbert space using maximum mean discrepancy [48]. Zhou *et al.* constructed a sparse feature transformation matrix based on compressive sensing theory to

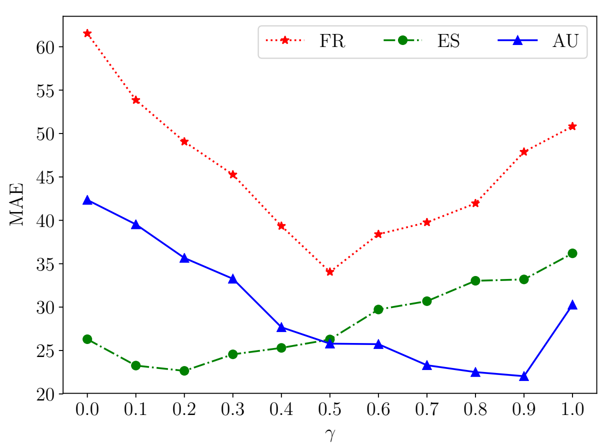

Figure 3: MAE under different γ values for the transfer learning models for FR, ES, and AU ($k = 1$).

map the weight vector of classifiers learned from the source domain to the target domain [70]. However, their tasks are very different from the regression task studied in this paper. These models were not able to capture efficiently the time series structure in our data.

Finally, the topic of disease modelling, and in particular of ILI, from online user-generated content has been extensively studied in the literature. The vast majority of methods proposed supervised solutions, using social media or search engine data together with disease rates from an established health authority [15, 21, 32, 33, 35, 38, 50, 52, 53, 67]. A few unsupervised methods have also been attempted, but they showcased moderate accuracy in terms of correlation [31, 51]. Our approach is able to provide accurate estimates without using any ground truth in the target locations.

6 CONCLUSIONS

Prior work on estimating disease rates from online user-generated content relies heavily on supervised learning models. Such models require ground truth data which is usually provided by public health organizations. Syndromic surveillance data, however, is either sparse or absent from locations with a poor healthcare infrastructure. This is somewhat ironic as it is often stated that web-based approaches hold considerable promise for regions that lack an established health surveillance system. This paper proposes a transfer learning framework as a potential solution to this problem. We leverage semantic and temporal relationships to map a supervised model from a source to a target location. We show that we can obtain a satisfactory performance ($r > .92$ on average) that does not deviate much from a fully supervised model ($\leq 21.6\%$ increase in RMSE), without using any ground truth from the target domain.

There is a number of avenues for future work. It is highly desirable to perform a study where the target country is from a low or middle income region. However, such a study is complicated, since the lack of ground truth data does not allow the performance to be quantified. Nevertheless, a qualitative study that demonstrated ILI estimates that followed an expected seasonal pattern would be of value. Our experiments on regions with ground truth data allowed us to investigate parameters k and γ, i.e. the choice for the one-to-k mapping and the relative weight assigned to the semantic and temporal similarities. Our analysis indicated that a one-to-one ($k = 1$) mapping performed best on average, and that the optimal γ differed per target country. Although we attempted to justify both outcomes, further experiments on other regions are needed to understand the effect of these parameters better.

ACKNOWLEDGMENTS

This work has been supported by the EPSRC grants EP/K031953/1 and EP/R00529X/1. The authors would like to acknowledge RCGP, PHE and SISSS for providing anonymized and aggregated syndromic surveillance data. Special thanks to Amparo Larrauri (SISSS) and Richard Pebody (PHE). We also thank Simon Moura and Peter Hayes for providing constructive feedback.

REFERENCES

[1] C. Alicino, N. L. Bragazzi, V. Faccio, D. Amicizia, D. Panatto, R. Gasparini, G. Icardi, and A. Orsi. 2015. Assessing Ebola-related Web Search Behaviour: Insights and Implications from an Analytical Study of Google Trends-based Query Volumes. *Infectious Diseases of Poverty* 4, 54 (2015), 1–13.

[2] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith. 2016. Massively Multilingual Word Embeddings. *arXiv Preprint* (2016), arXiv:1602.01925.

[3] H. K. Andreassen, M. M. Bujnowska-Fedak, C. E. Chronaki, R. C. Dumitru, I. Pudule, S. Santana, H. Voss, and R. Wynn. 2007. European Citizens' Use of E-health Services: A Study of Seven Countries. *BMC Public Health* 7, 53 (2007).

[4] M. Artetxe, G. Labaka, and E. Agirre. 2016. Learning Principled Bilingual Mappings of Word Embeddings while Preserving Monolingual Invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2289–2294.

[5] M. Artetxe, G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised Neural Machine Translation. In *Proceedings of the 6th International Conference on Learning Representations.*

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. 2010. A Theory of Learning from Different Domains. *Machine Learning* 79, 1-2 (2010), 151–175.

[7] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. 2007. Analysis of Representations for Domain Adaptation. In *Proceedings of Advances in Neural Information Processing Systems 19.* 137–144.

[8] A. Benton, R. Arora, and M. Dredze. 2016. Learning Multiview Embeddings of Twitter Users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 14–19.

[9] M. Biggerstaff, M. Johansson, D. Alper, L. C. Brooks, P. Chakraborty, D. C. Farrow, S. Hyun, S. Kandula, C. McGowan, N. Ramakrishnan, et al. 2018. Results from the Second Year of A Collaborative Effort to Forecast Influenza Seasons in the United States. *Epidemics* 24 (2018), 26–33.

[10] J. Blitzer, R. McDonald, and F. Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.* 120–128.

[11] G. Boivin, I. Hardy, G. Tellier, and J. Maziade. 2000. Predicting Influenza Infections during Epidemics with Use of a Clinical Case Definition. *Clinical Infectious Diseases* 31, 5 (2000), 1166–1169.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association of Computational Linguistics* 5, 1 (2017), 135–146.

[13] R. Caruana. 1998. Multitask Learning. In *Learning to Learn.* Springer, 95–133.

[14] M. Chen, K. Q. Weinberger, and J. Blitzer. 2011. Co-Training for Domain Adaptation. In *Proceedings of Advances in Neural Information Processing Systems 24.* 2456–2464.

[15] A. Culotta. 2010. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proceedings of the 1st Workshop on Social Media Analytics.* 115–122.

[16] W. Dai, G. Xue, Q. Yang, and Y. Yu. 2007. Transferring Naive Bayes Classifiers for Text Classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence.* 540–545.

[17] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media.* 128–137.

[18] G. Dinu, A. Lazaridou, and M. Baroni. 2014. Improving Zero-shot Learning by Mitigating the Hubness Problem. *arXiv Preprint* (2014), arXiv:1412.6568.

[19] L. Duan, D. Xu, and I. W. Tsang. 2012. Learning with Augmented Features for Heterogeneous Domain Adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning.* 667–674.

[20] G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* 451–459.

[21] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. Detecting Influenza Epidemics using Search Engine Query Data. *Nature* 457, 7232 (2009), 1012–1014.

[22] X. Glorot, A. Bordes, and Y. Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning.* 513–520.

[23] G. H. Golub and C. Reinsch. 1970. Singular Value Decomposition and Least Squares Solutions. *Numerische Mathematik* 14, 5 (1970), 403–420.

[24] J. Gomide, A. Veloso, W. M. Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. 2011. Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. In *Proceedings of the 3rd International Web Science Conference.* Article 3, 3:1–3:8 pages.

[25] O. Higgins, J. Sixsmith, M. M. Barry, and C. Domegan. 2011. *A Literature Review on Health Information Seeking Behaviour on the Web: A Health Consumer and Health Professional Perspective.* Technical Report.

[26] A. E. Hoerl and R. W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12 (1970), 55–67.

[27] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. 2013. Cross-Language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* 7304–7308.

[28] J. Jiang and C. Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* 264–271.

[29] P. Koehn and J. Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation.* 224–227.

[30] B. Kulis, K. Saenko, and T. Darrell. 2011. What you Saw is not What you Get: Domain Adaptation using Asymmetric Kernel Transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1785–1792.

[31] A. Lamb, M. J. Paul, and M. Dredze. 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 789–795.

[32] V. Lampos and N. Cristianini. 2010. Tracking the flu pandemic by monitoring the social Web. In *Proceedings of the 2nd International Workshop on Cognitive Information Processing.* IEEE Press, 411–416.

[33] V. Lampos and N. Cristianini. 2012. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology* 3, 4 (2012), 1–22.

[34] V. Lampos, T. De Bie, and N. Cristianini. 2010. Flu Detector – Tracking Epidemics on Twitter. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III (ECML PKDD'10).* Springer, 599–602.

[35] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports* 5, 12760 (2015).

[36] V. Lampos, D. Preoţiuc-Pietro, and T. Cohn. 2013. A user-centric model of voting intention from Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 993–1003.

[37] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. 2015. Assessing the Impact of a Health Intervention via User-Generated Internet Content. *Data Mining and Knowledge Discovery* 29, 5 (2015), 1434–1457.

[38] V. Lampos, B. Zou, and I. J. Cox. 2017. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In *Proceedings of the 26th International Conference on World Wide Web.* 695–704.

[39] O. Levy, A. Søgaard, and Y. Goldberg. 2017. A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1. 765–774.

[40] Y. Mansour, M. Mohri, and A. Rostamizadeh. 2009. Domain Adaptation: Learning Bounds and Algorithms. *arXiv Preprint* (2009), arXiv:0902.3430.

[41] T. Mikolov, Q. V. Le, and I. Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *arXiv Preprint* (2013), arXiv:1309.4168.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26.* 3111–3119.

[43] G. J Milinovich, G. M Williams, A. C. A. Clements, and W. Hu. 2014. Internet-based Surveillance Systems for Monitoring Emerging Infectious Diseases. *The Lancet Infectious Diseases* 14, 2 (2014), 160–168.

[44] A. Mogadala and A. Rettinger. 2016. Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-language Text Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 692–702.

[45] M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar. 2011. Google Correlate Whitepaper. https://www.google.com/trends/correlate/whitepaper.pdf.

[46] A. S. Monto, S. Gravenstein, M. Elliott, M. Colopy, and J. Schweinle. 2000. Clinical Signs and Symptoms Predicting Influenza Infection. *Archives of Internet Medicine* 160, 21 (2000), 3243–3247.

[47] N. Mrkšić, I. Vulić, D. Ó Séaghdha, I. Leviant, R. Reichart, M. Gašić, A. Korhonen, and S. Young. 2017. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-lingual Constraints. *arXiv Preprint* (2017),

arXiv:1706.00374.
[48] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. 2009. Domain Adaptation via Transfer Component Analysis. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. 1187–1192.
[49] S. J. Pan and Q. Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.
[50] M. J. Paul and M. Dredze. 2011. You Are What You Tweet: Analysing Twitter for Public Health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. 265–272.
[51] Michael J. Paul and Mark Dredze. 2014. Discovering Health Topics in Social Media Using Topic Models. *PLOS One* 9, 8 (2014), e103408.
[52] M. J. Paul, M. Dredze, and D. Broniatowski. 2014. Twitter Improves Influenza Forecasting. *PLOS Currents* 6 (2014).
[53] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. 2008. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases* 47, 11 (2008), 1443–1448.
[54] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. 2015. Studying User Income through Language, Behaviour and Affect in Social Media. *PLOS ONE* 10, 9 (2015), e0138717.
[55] P. Prettenhofer and B. Stein. 2010. Cross-language Text Classification using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 1118–1127.
[56] S. Ruder. 2017. A Survey of Cross-Lingual Embedding Models. *arXiv Preprint* (2017), arXiv:1706.04902.
[57] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. 2016. Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. *arXiv Preprint* (2016), arXiv:1702.03859.
[58] R. Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58, 1 (1996), 267–288.
[59] L. Torrey and J. Shavlik. 2009. Transfer Learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques* (2009), 242.
[60] I. Vulić and M.-F. Moens. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2. 719–725.
[61] I. Vulić and M.-F. Moens. 2015. Monolingual and Cross-lingual Information Retrieval Models based on (Bilingual) Word Embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 363–372.
[62] I. Vulić and M.-F. Moens. 2016. Bilingual Distributed Word Representations from Document-Aligned Comparable Data. *Journal of Artificial Intelligence Research* 55 (2016), 953–994.
[63] M. Wagner, V. Lampos, I. J. Cox, and R. Pebody. 2018. The Added Value of Online User-generated Content in Traditional Methods for Influenza Surveillance. *Scientific Reports* 8, 13963 (2018).
[64] X. Wan. 2009. Co-training for Cross-lingual Sentiment Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association of Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. 235–243.
[65] K. Weiss, T. M. Khoshgoftaar, and D. Wang. 2016. A Survey of Transfer Learning. *Journal of Big Data* 3, 1 (2016).
[66] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao. 2015. Short Text Clustering via Convolutional Neural Networks.. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 62–69.
[67] S. Yang, M. Santillana, and S. C. Kou. 2015. Accurate Estimation of Influenza Epidemics using Google Search Data via ARGO. *Proceedings of the National Academy of Sciences* 112, 47 (2015), 14473–14478.
[68] M. Ybarra and M. Suman. 2008. Reasons, Assessments and Actions Taken: Sex and Age Differences in Uses of Internet Health Information. *Health Education Research* 23, 3 (2008), 512–521.
[69] P. Zhao and B. Yu. 2006. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* 7 (2006), 2541–2563.
[70] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan. 2014. Heterogeneous Domain Adaptation for Multiple Classes. In *Artificial Intelligence and Statistics*. 1095–1103.
[71] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. 2011. Heterogeneous Transfer Learning for Image Classification. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. 1304–1309.
[72] B. Zou, V. Lampos, and I. J. Cox. 2018. Multi-Task Learning Improves Disease Models from Web Search. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 87–96.
[73] B. Zou, V. Lampos, R. Gorton, and I. J. Cox. 2016. On Infectious Intestinal Disease Surveillance using Social Media Content. In *Proceedings of the 6th International Conference on Digital Health*. ACM, 157–161.
[74] H. Zou and T. Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (2005), 301–320.
[75] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. 2013. Bilingual Word Embeddings for Phrase-based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1393–1398.