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

Zou, Bin; Lampos, Vasileios; Cox, Ingemar

Published in:
WWW '19: The World Wide Web Conference

DOI:
10.1145/3308558.3313477

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data

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 290,000 deaths worldwide on an annual basis.1 Data from the microblogging platform of Twitter [15, 32, 50] as well as from search engines [21, 29, 42] 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.

1World Health Organization, who.int/mediacentre/news/statements/2017/flu/
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 (≤ 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), GP’s Sentinelles Network (SN), Spanish Influenza Sentinel Surveillance System (SISSS), and Australian Sentinel Practices Research Network (ASPREN), respectively. ILI rates represent fraction of the population that has been diagnosed with influenza-like symptoms. 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. 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 insights 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 ≈ 3 × 10⁻²⁴), 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 ≈ 8 × 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 \( f : X \rightarrow y \) that maps the input space of search query frequencies, \( X \in \mathbb{R}^{n \times s} \), to the target variable, \( y \in \mathbb{R}^m \), 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, \( X \) contains the time series of search query frequencies, and \( y \) 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 \( y \) (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, \( f : X \rightarrow y \), 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 \( \mathcal{D}_S = \{(x_i, y_i)\} \).
\( i \in \{1, \ldots, n\} \), 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 \( D_T = \{ \mathbf{x}_i' \}, i \in \{1, \ldots, m\} \), 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 use 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 \( \rightarrow \) gripe (FR) \( \rightarrow \) gripe (ES)).

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 \( \rightarrow \) gripe (FR) \( \rightarrow \) gripe (ES)). 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 \( \rightarrow \) gripe (FR) \( \rightarrow \) gripe (ES)). 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 \( \rightarrow \) gripe (FR) \( \rightarrow \) gripe (ES)).

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 \( \ell_1 \)-norm regularization, commonly known as the lasso [58], with \( \ell_2 \)-norm, or ridge [26], regularization. In addition to the sparsity encouraged by the \( \ell_1 \)-norm regularization, the \( \ell_2 \)-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 \( X \in \mathbb{R}^{n \times s} \) and \( y \in \mathbb{R}^n \) from the source domain \( D_S \), we apply a constrained version of elastic net which solves the following optimization problem:

\[
\arg\min_{w, \beta} \left\{ \|y - Xw - \beta\|^2_2 + \lambda_1 \|w\|^2_1 + \lambda_2 \|w\|_2 \right\} \quad \text{subject to} \quad w \geq 0, \tag{1}
\]

where \( \lambda_1 > 0 \) and \( \lambda_2 > 0 \) are respectively the \( \ell_1 \)-norm and \( \ell_2 \)-norm regularization parameters, and \( \beta \) denotes the intercept term. The non-negativity constraint for \( w \) 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, \( T \), was defined as a simple set of two flu-related terms, \( T = \{\text{flu}, \text{`fever'}\} \). 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 \( T \) and \( e_q \), i.e.

\[
g(q, T) = \cos\left(e_q, e_{f_1}\right) \times \cos\left(e_q, e_{f_2}\right), \tag{2}
\]

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

This resolves misleading similarity scores based on different sign combinations.
We define a combined similarity metric, \( \Theta \) by exploiting sources of monolingual text alongside a smaller cross-lingual text. The majority of cross-lingual word embedding models are trained by exploiting sources of monolingual text alongside a smaller cross-lingual text. Empirical evidence indicates that they can also embed cross-lingual representations for 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 \([57, 60, 61]\).\(^4\) The dimensionality of the word embeddings was set to \( d = 300 \). Then, we used \( \alpha \) a core selection of exact translation pairs \((\alpha \rightarrow \tau)\) from the source to the target domain language to generate bilingual embeddings. Given the embedding matrices of this alignment dictionary, \( E_S \) and \( E_T \), 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 \( W \in \mathbb{R}^{d \times d} \) such that \( E_T = E_S W \). An orthogonal matrix learned by minimizing the squared Euclidean distance between \( E_S \) and \( E_T \), i.e.

\[
\arg\min_W \| E_S W - E_T \|_2^2, \quad \text{subject to } W^T W = I. \tag{4}
\]

The orthogonality constraint ensures that the transformation works both ways, that is \( E_T \approx E_S W \), \( E_S \approx E_T W \), and \( E_T = E_S W \) \([57]\). In addition, Arthe et al. have empirically shown that it also improves the performance of machine translation \([4]\). The exact solution of Eq. 4 is given by \( W = UV^\top \), where \( E_T^\top E_S = U \Sigma V^\top \) is the singular value decomposition of \( E_T^\top E_S \) \([4, 23]\).

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 \( \|v \|_2 \) the Euclidean distance between \( v \), the embeddings of a source query and \( \|v\|_2 \) the embeddings of a target query. The cosine similarity between \( v \) and \( w \), the embeddings of \( N \) source and valid target queries is given by \( \omega_{ij} = \langle v_i, v_j \rangle / (\|v_i\|_2 \|v_j\|_2) \). Note that the cosine similarities are computed after projecting the embeddings of the source domain to the target domain using the transformation matrix \( W \).

In theory, we can directly use \( \omega_{ij} \) 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

\[
P_{j \rightarrow i} = \frac{\exp{\left( \eta \omega_{ij} \right)}}{\alpha_j \sum_{z=1}^{N} \exp{\left( \eta \omega_{iz} \right)}},
\]

where \( \alpha_j \) is a normalization factor that ensures \( P_{j \rightarrow i} \) is a probability, and \( s \) is the number of source queries in the vocabulary.

\[^3\]The embeddings were obtained from \url{github.com/facebookresearch/fastText}.

\[^4\]We have empirical evidence about this, obtained during the first stages of this work.

---

<table>
<thead>
<tr>
<th>Search queries</th>
<th>US</th>
<th>FR</th>
<th>ES</th>
<th>AU</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu (US/AU), gripe (FR), gripe (ES)</td>
<td>0.036 (.010)</td>
<td>0.033 (.012)</td>
<td>0.032 (.011)</td>
<td>0.031 (.016)</td>
</tr>
<tr>
<td>symptoms of flu (US/AU), symptômes de la grippe (FR), síntomas de gripe (ES)</td>
<td>0.030 (.009)</td>
<td>0.031 (.012)</td>
<td>0.029 (.009)</td>
<td>0.027 (.014)</td>
</tr>
<tr>
<td>flu in children (US/AU), gripe chez le bébé (FR), gripe en el bebé (ES)</td>
<td>0.017 (.007)</td>
<td>0.020 (.008)</td>
<td>0.019 (.009)</td>
<td>0.022 (.010)</td>
</tr>
</tbody>
</table>
the other way around, $P_i \rightarrow j$ [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 $\eta$ is learned by maximizing the log probability over the alignment dictionary ($\sigma \rightarrow \tau$), i.e., $\arg \max_{\eta} \sum_{i,j} \ln \left( P_{i \rightarrow j} \right)$. The top-$k$ queries from $\mathcal{P}_T$ with the highest pairing probability ($P_{i \rightarrow j}$) are then selected as possible translations of the source query $q_i$. Finally, we compute the semantic (cosine) similarity score $\Theta_k$ between the source query $q_i$ and the target query $q_j$ using $\Theta_k(q_i, q_j) = \frac{\langle \mathbf{e}_{q_i}, \mathbf{W} \mathbf{e}_{q_j}^T \rangle}{\|\mathbf{e}_{q_i}\| \|\mathbf{e}_{q_j}\|}$, 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, $Q_S$, may not be present in the pool of candidate target queries, $\mathcal{P}_T$. 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 ($\Theta_t$).** 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 ±$\xi$ weeks. The range of possible values for $\xi$ 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 $\mathcal{T}_k$ (consisting of $k \geq 1$ queries from $\mathcal{P}_T$), and their associated daily search frequencies, $x_i(t)$ and $x_j(t)$, respectively, the temporal correlation similarity, $\Theta_t$, is given by

$$\Theta_t(q_i, q_j) = \rho \left( x_i(t), x_j(t+\xi) \right),$$

where $\rho(x_i(t), x_j(t+\xi))$ denotes the optimal Pearson correlation coefficient between $x_i$, $x_j$ within the shifting window. Note that the optimal window is independently computed for each target query in $\mathcal{T}_k$ 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, $\mathcal{T}_k$, 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 $q_j \in \mathcal{T}_k$, and assign each query in $\mathcal{T}_k$ 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 $\Theta_{ij}$, $j \in \{2, \ldots, k\}$, with the source query (see Eq. 3). More specifically, a target weight $w_j'$ is defined as $w_j' = \frac{w_i \Theta_{ij}}{\sum_{q_j \in \mathcal{T}_k} \Theta_{ij}}$.

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. 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 ($\xi$) used for identifying optimal correlations between the frequency time series of the source and target queries (see Section 3) is set to ±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 ($\rho$), 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’,
### Table 2: Performance estimates for the US→FR transfer learning task. Different values of \( \gamma \) determine how queries are mapped from the source to the target domain (\( \gamma = 1 \): semantic similarity only, \( \gamma = 0 \): temporal correlation only, \( \gamma = 0.1 \): 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 \( \gamma \) value is underlined. Only the best random mapping performance (R) is enumerated per choice of \( \gamma \). The last two rows show the performance of the baseline models.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MAE RMSE</td>
<td>MAE RMSE</td>
<td>MAE RMSE</td>
<td>MAE RMSE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>1</td>
<td>—</td>
<td>797 7.98</td>
<td>136.999</td>
<td>789 9.584</td>
<td>139.752</td>
<td>900 36.107</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>—</td>
<td>803 80.04</td>
<td>137.247</td>
<td>794 9.961</td>
<td>148.835</td>
<td>890 58.372</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>—</td>
<td>792 9.016</td>
<td>135.995</td>
<td>796 9.759</td>
<td>139.350</td>
<td>896 57.241</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>4</td>
<td>798 9.077</td>
<td>135.892</td>
<td>795 9.529</td>
<td>95.295</td>
<td>895 58.380</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>5</td>
<td>799 8.881</td>
<td>135.743</td>
<td>798 5.808</td>
<td>95.036</td>
<td>893 58.439</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>2</td>
<td>1</td>
<td>803 8.012</td>
<td>137.170</td>
<td>794 9.971</td>
<td>148.869</td>
<td>890 58.360</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>2</td>
<td>3</td>
<td>802 7.998</td>
<td>135.885</td>
<td>796 9.763</td>
<td>95.360</td>
<td>896 57.444</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>3</td>
<td>4</td>
<td>799 7.968</td>
<td>135.725</td>
<td>795 9.529</td>
<td>95.373</td>
<td>895 58.365</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>1</td>
<td>1</td>
<td>771 12.452</td>
<td>152.275</td>
<td>731 9.122</td>
<td>105.769</td>
<td>805 25.102</td>
</tr>
</tbody>
</table>

**Table Notes:**
- \( k \): number of target queries (1-to-1 mapping).

**Notes:**
- \( \gamma = 0 \): pure semantic similarity.
- \( \gamma = 0 \): pure temporal correlation.
- \( \gamma = 0.1 \): joint semantic and temporal correlation.

**Discussion:**

- The performance estimates indicate that the joint correlation-based approach outperforms the purely semantic and temporal approaches, especially when \( \gamma = 0.1 \).
- The non-uniform weighting scheme improves performance, especially for \( \gamma = 0 \), suggesting that it better captures the variability in the data.
- The baseline models (R) consistently outperform the transfer learning models, indicating that the default mapping is highly effective.

---

<table>
<thead>
<tr>
<th>( \gamma = 1 )</th>
<th>( \gamma_{opt} = .5 )</th>
<th>Unsupervised</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k: 1</td>
<td>964</td>
<td>31.558</td>
<td>41.675</td>
</tr>
<tr>
<td>w: 1</td>
<td>963</td>
<td>32.302</td>
<td>42.384</td>
</tr>
<tr>
<td>09/2012 – 09/2013</td>
<td>964</td>
<td>31.558</td>
<td>41.675</td>
</tr>
<tr>
<td>09/2013 – 09/2014</td>
<td>965</td>
<td>32.081</td>
<td>42.293</td>
</tr>
<tr>
<td>09/2014 – 09/2015</td>
<td>964</td>
<td>31.918</td>
<td>41.791</td>
</tr>
<tr>
<td>09/2015 – 09/2016</td>
<td>963</td>
<td>32.302</td>
<td>42.384</td>
</tr>
<tr>
<td>Average</td>
<td>963</td>
<td>32.302</td>
<td>42.384</td>
</tr>
</tbody>
</table>

**Notes:**
- For \( \gamma = 1 \), we observe a significant drop in performance for \( \gamma = 0 \).
- The optimal choice of \( \gamma \) is \( \gamma_{opt} = .5 \) for both supervised and unsupervised learning.
- The unsupervised learning approach, while generally inferior, shows improved performance with \( \gamma = .5 \), especially in the early stages of learning.

---

**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, \( P_T \), 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

---

**Conclusion:**

The joint correlation-based weighting scheme significantly outperforms both purely semantic and purely temporal approaches, demonstrating the importance of considering the temporal aspect in transfer learning tasks.

---

[Note: The text in the image is not fully legible due to the quality of the scan. The above transcription is an attempt to accurately represent the content as best possible.]
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 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 γ = 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 γ = .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 diverge much from the default settings of γ = .5 and k = 1, there are discrepancies between the optimal γ value for each task (\(γ_{opt} = .5, .2\) 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 TC, which performs even worse than random (results are omitted). One explanation for this is that TC 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 one 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 3: Performance estimates for US→ES transfer learning task. Please refer to Table 2’s caption for further information.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>r MAE</td>
<td>RMSE</td>
<td>r MAE</td>
<td>RMSE</td>
<td>r MAE</td>
<td>RMSE</td>
<td>r MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>-------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>---------</td>
</tr>
<tr>
<td>1</td>
<td>808</td>
<td>25.068</td>
<td>41.104</td>
<td>807</td>
<td>25.789</td>
<td>42.137</td>
<td>843</td>
</tr>
<tr>
<td>2</td>
<td>799</td>
<td>25.589</td>
<td>42.631</td>
<td>835</td>
<td>23.850</td>
<td>39.092</td>
<td>844</td>
</tr>
<tr>
<td>3</td>
<td>795</td>
<td>25.756</td>
<td>42.883</td>
<td>830</td>
<td>24.589</td>
<td>37.913</td>
<td>831</td>
</tr>
<tr>
<td>4</td>
<td>783</td>
<td>26.504</td>
<td>43.662</td>
<td>835</td>
<td>23.671</td>
<td>39.207</td>
<td>844</td>
</tr>
<tr>
<td>6</td>
<td>799</td>
<td>25.579</td>
<td>42.610</td>
<td>835</td>
<td>23.850</td>
<td>39.095</td>
<td>844</td>
</tr>
<tr>
<td>7</td>
<td>795</td>
<td>25.748</td>
<td>42.867</td>
<td>835</td>
<td>23.670</td>
<td>38.932</td>
<td>842</td>
</tr>
<tr>
<td>8</td>
<td>794</td>
<td>25.641</td>
<td>42.633</td>
<td>835</td>
<td>23.671</td>
<td>39.209</td>
<td>842</td>
</tr>
<tr>
<td>9</td>
<td>830</td>
<td>40.548</td>
<td>46.584</td>
<td>903</td>
<td>36.241</td>
<td>47.781</td>
<td>866</td>
</tr>
</tbody>
</table>

\(k\): number of target queries (1-to-k mapping), \(w\): weighting approach, U: uniform, NU: non-uniform, C: correlation, R: random.
Table 4: Performance estimates for the US→AU transfer learning task. Please refer to Table 2’s caption for further information.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>r MAE  RMSE</td>
<td>r MAE  RMSE</td>
<td>r MAE  RMSE</td>
<td>r MAE  RMSE</td>
<td>r MAE   RMSE</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>---------</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>704</td>
<td>38.804 50.140</td>
<td>679</td>
<td>39.708 48.508</td>
<td>630</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>622</td>
<td>41.824 55.943</td>
<td>663</td>
<td>41.708 50.752</td>
<td>633</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>621</td>
<td>42.263 56.819</td>
<td>669</td>
<td>42.900 51.468</td>
<td>679</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>607</td>
<td>42.040 56.755</td>
<td>669</td>
<td>42.501 51.008</td>
<td>634</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>600</td>
<td>41.900 56.618</td>
<td>671</td>
<td>41.950 49.692</td>
<td>647</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>623</td>
<td>41.886 55.947</td>
<td>663</td>
<td>41.642 50.818</td>
<td>633</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>620</td>
<td>42.263 56.812</td>
<td>669</td>
<td>42.857 51.533</td>
<td>631</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>607</td>
<td>42.031 56.745</td>
<td>669</td>
<td>42.466 51.039</td>
<td>634</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>600</td>
<td>41.885 56.601</td>
<td>671</td>
<td>41.928 49.723</td>
<td>647</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>653</td>
<td>60.835 71.392</td>
<td>710</td>
<td>50.090 62.045</td>
<td>628</td>
</tr>
</tbody>
</table>

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

4.2 Qualitative analysis

A fair criticism for the proposed framework is that in a practical scenario the optimal values for $k$ and $c$ cannot be validated. However, we have already demonstrated that the default settings of $k = .5$ 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 $k$. As discussed previously, optimal $y$ values differ per target country. Interestingly, all error trends are monotonically decreasing (as $y$ increases) until they reach a minimum, and then begin to monotonically increase. We argue that $y_{opt}$ reflects on the actual pool of candidate target queries ($P_T$), 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 $y$ values (.5, .2, and .9), an insight that can be used to make a more informed choice of $y$ 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 $\times$ 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%), “chill fever” to “sifnosis” (7.7%), and “child temperature” to “warmer” (9.8%). Nevertheless, it is encouraging.

$^{9}$“Grippe intestinale” translates to “stomach flu” (formally “viral gastroenteritis”).
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

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.
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, Prentboer 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 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 $\gamma$, 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 $\gamma$ 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, RGP and SISSS for providing anonymized and aggregated syndromic surveillance data. Special thanks to Amparo Larrauri (SISSS) and Richard Pehoe (PHE). We also thank Simon Moura and Peter Hayes for providing constructive feedback.

REFERENCES


