Transfer learning for unsupervised influenza-like illness models from online search data

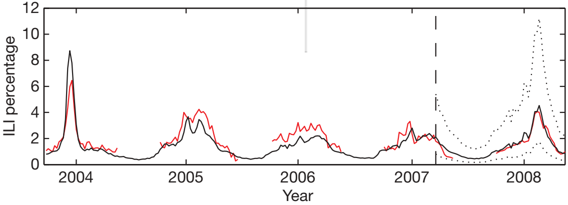
Bin Zou
Vasileios Lampos (`lampos.net`)
Ingemar J. Cox

Department of Computer Science
University College London

From online searches to influenza-like illness rates

flu treatment

- flu treatment
- flu treatment **kids**
- flu treatment **otc**
- flu treatment **natural**
- flu treatment **medication**
- flu treatment **toddler**



From online searches to influenza-like illness rates

Google Flu Trends (*discontinued*) Ginsberg et al. (2009)

popularising an established idea Eysenbach (2006); Polgreen et al. (2008)

From online searches to influenza-like illness rates

Task abstraction

- **input** – frequency of search queries over time: $\mathbf{X} \in \mathbb{R}^{n \times s}$
- **output** – corresponding influenza-like illness (ILI) rate: $\mathbf{y} \in \mathbb{R}^{n}$
- **regression task**, i.e. learn $f : \mathbf{X} \rightarrow \mathbf{y}$

From online searches to influenza-like illness rates

Task abstraction

- **input** – frequency of search queries over time: $\mathbf{X} \in \mathbb{R}^{n \times s}$
- **output** – corresponding influenza-like illness (ILI) rate: $\mathbf{y} \in \mathbb{R}^{n}$
- **regression task**, i.e. learn $f : \mathbf{X} \rightarrow \mathbf{y}$

Modelling

- originally proposed models were evidently **not good solutions**[1]
- **new families of methods** seem to work OK in various geographies[2]

[1] Cook et al. (2011); Olson et al. (2013); Lazer et al. (2014)

[2] Lampos et al. (2015a); Yang et al. (2015); Lampos et al. (2017); Wagner et al. (2018)

Why estimate ILI rates from online search statistics?

Common arguments for:

- complements traditional syndromic surveillance
  - ✓ timeliness
  - ✓ broader demographic coverage, larger cohort
  - ✓ broader geographical coverage
  - ✓ not affected by closure days or national holidays
  - ✓ lower cost
- applicable to locations that **lack an established health system**

Why estimate ILI rates from online search statistics?

Common arguments for:

- complements traditional syndromic surveillance
  - ✓ timeliness
  - ✓ broader demographic coverage, larger cohort
  - ✓ broader geographical coverage
  - ✓ not affected by closure days or national holidays
  - ✓ lower cost
- applicable to locations that **lack an established health system**
  - ✓ **oxymoron** (*supervised learning*)

Why estimate ILI rates from online search statistics?

Common arguments for:

- complements traditional syndromic surveillance
  - ✓ timeliness
  - ✓ broader demographic coverage, larger cohort
  - ✓ broader geographical coverage
  - ✓ not affected by closure days or national holidays
  - ✓ lower cost
- applicable to locations that **lack an established health system**
  - ✓ **oxymoron** (*supervised learning*)
  - ✓ **motivated this paper**

Our contribution in a nutshell

Main task

- train a model for a **source location** where historical syndromic surveillance data is available, and
- transfer it to a **target location** where syndromic surveillance data is not available or, in our experiments, ignored

Our contribution in a nutshell

Main task

- train a model for a **source location** where historical syndromic surveillance data is available, and
- transfer it to a **target location** where syndromic surveillance data is not available or, in our experiments, ignored

Transfer learning steps

1. Learn a **linear regularised regression model** for a **source** location
2. **Map search queries** from the source to the target domain (languages may differ)
3. **Transfer the source weights** to the target domain (might involve weight re-adjustment)

Transfer learning task definition

$$\text{query frequency } x_{ij} = \frac{\#\text{query } j \text{ issued during } \Delta t_i}{\#\text{all queries issued during } \Delta t_i} \text{ for a location}$$

Source domain

- $\mathcal{D}_\mathsf{S} = \left\{ \, (\mathbf{x}_i, y_i) \, \right\}$, $i \in \{1, ..., n\}$
- $\mathbf{x}_i \in \mathbb{R}^s = \{x_{ij}\}$, $j \in \{1, ..., s\}$: frequency of source queries
- $y_i \in \mathbb{R}$: ILI rate for time interval i

Target domain

- $\mathcal{D}_\mathsf{T} = \{\mathbf{x}'_i\}$, $i \in \{1, ..., m\}$
- $\mathbf{x}'_i \in \mathbb{R}^t$: frequency of target queries
- note that t need not equal s

Transfer learning task definition

$$\text{query frequency } x_{ij} = \frac{\#\text{query } j \text{ issued during } \Delta t_i}{\#\text{all queries issued during } \Delta t_i} \text{ for a location}$$

Source domain

- $\mathcal{D}_\mathsf{S} = \left\{ (\mathbf{x}_i, y_i) \right\}$, $i \in \{1, \ldots, n\}$
- $\mathbf{x}_i \in \mathbb{R}^s = \{x_{ij}\}$, $j \in \{1, \ldots, s\}$: frequency of source queries
- $y_i \in \mathbb{R}$: ILI rate for time interval i

Target domain

- $\mathcal{D}_\mathsf{T} = \{\mathbf{x}'_i\}$, $i \in \{1, \ldots, m\}$
- $\mathbf{x}'_i \in \mathbb{R}^t$: frequency of target queries
- note that t need not equal s

Aim: Given $\mathcal{D}_\mathsf{S}$ and $\mathcal{D}_\mathsf{T}$, estimate y'_i

Step 1 – Learn a regression function in the source domain

Source domain

- $\mathbf{x}_i \in \mathbb{R}^s = \{x_{ij}\}, j \in \{1, \ldots, s\}$: frequency of source queries
- $y_i \in \mathbb{R}$: ILI rate for time interval i

Elastic net[1] (*constrained*)

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta - \left(\sum_{j=1}^{s} x_{ij} w_j \right) \right)^2 + \lambda_1 \sum_{j=1}^{s} |w_j| + \lambda_2 \sum_{j=1}^{s} w_j^2$$

subject to $\mathbf{w} \geq 0$

[1] Zou and Hastie (2005)

Step 1 – Learn a regression function in the source domain

Elastic net (*constrained*)

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta - \left(\sum_{j=1}^{s} x_{ij} w_j \right) \right)^2 + \lambda_1 \sum_{j=1}^{s} |w_j| + \lambda_2 \sum_{j=1}^{s} w_j^2$$

subject to $\mathbf{w} \geq 0$

Why use elastic net?

- more straightforward to transfer
- few training instances
- previous successful application[1]
- combines ℓ_1- and ℓ_2-norm regularisation: sparse solution, model consistency under collinearity

[1] Lampos et al. (2015a,b); Zou et al. (2016); Lampos et al. (2017)

Step 1 – Learn a regression function in the source domain

Elastic net (*constrained*)

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta - \left(\sum_{j=1}^{s} x_{ij} w_j \right) \right)^2 + \lambda_1 \sum_{j=1}^{s} |w_j| + \lambda_2 \sum_{j=1}^{s} w_j^2$$

subject to $\mathbf{w} \geq 0$

Why apply a non-negative weight constraint?

- (*how?*) coordinate descent restricting negative updates to 0
- worse performing model for the source location
- **but** enables a more comprehensive transfer
- better performance at the target location

Step 1 – Learn a regression function in the source domain

Selecting queries prior to applying elastic net

- **hybrid feature selection** similarly to previous work[1]
- derive query embeddings $\mathbf{e}_{\mathsf{q}}$ using `fastText`[2]
- define a flu context/topic: $\mathcal{T} = \{$‘*flu*’, ‘*fever*’$\}$
- compute each query's similarity to $\mathcal{T}$ using

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

$\cos(\cdot,\cdot)$ is mapped to $[0, 1]$

[1] Zou et al. (2016); Lampos et al. (2017); Zou et al. (2018)

[2] Bojanowski et al. (2017)

Step 1 – Learn a regression function in the source domain

Selecting queries prior to applying elastic net

- **hybrid feature selection** similarly to previous work[1]
- derive query embeddings $\mathbf{e}_{\mathsf{q}}$ using `fastText`[2]
- define a flu context/topic: $\mathcal{T} = \{$'*flu*', '*fever*'$\}$
- compute each query's similarity to $\mathcal{T}$ using

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

$\cos(\cdot, \cdot)$ is mapped to $[0, 1]$

- filter out queries with either $g \leq 0.5$ **or** $r \leq 0.3$ (corr. with ILI)

$\mathcal{Q}_{\mathsf{S}}$: remaining queries after applying elastic net

[1] Zou et al. (2016); Lampos et al. (2017); Zou et al. (2018)

[2] Bojanowski et al. (2017)

Step 2 – Mapping source to target queries

Task: map $\mathcal{Q}_\mathsf{S}$ to a subset of $\mathcal{P}_\mathsf{T}$ (pool of target queries)

Step 2 – Mapping source to target queries

Task: map $\mathcal{Q}_\mathsf{S}$ to a subset of $\mathcal{P}_\mathsf{T}$ (pool of target queries)

How?

- direct translation **does not work**
  — invalid search queries
  — worse performance

Step 2 – Mapping source to target queries

Task: map $\mathcal{Q}_\mathsf{S}$ to a subset of $\mathcal{P}_\mathsf{T}$ (pool of target queries)

How?

- direct translation **does not work**
  — invalid search queries
  — worse performance
- **semantic similarity**, Θ_s**:** (*cross-lingual*) word embeddings
- **temporal similarity**, Θ_c**:** correlation between frequency time series
- **hybrid similarity:** $\Theta = \gamma\Theta_\mathsf{s} + (1-\gamma)\Theta_\mathsf{c}$, $\gamma \in [0, 1]$
- consider 1-to-k mappings

Step 2 – Semantic similarity (Θ_s)

Same language in both domains?

- Use cosine similarity on query embeddings

Step 2 – Semantic similarity (Θ_s)

Same language in both domains?

- Use cosine similarity on query embeddings

If not, **derive bi-lingual embeddings**[1]

- m core translation pairs, $\sigma \rightarrow \tau$, with embeddings $\mathbf{E}_\sigma$, $\mathbf{E}_\tau \in \mathbb{R}^{m \times d}$
- learn a transformation matrix, $\mathbf{W} \in \mathbb{R}^{d \times d}$, by minimising:

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

[1] Smith et al. (2016)

Step 2 – Semantic similarity (Θ_s)

Same language in both domains?

- Use cosine similarity on query embeddings

If not, **derive bi-lingual embeddings**[1]

- m core translation pairs, $\sigma \rightarrow \tau$, with embeddings $\mathbf{E}_\sigma$, $\mathbf{E}_\tau \in \mathbb{R}^{m \times d}$
- learn a transformation matrix, $\mathbf{W} \in \mathbb{R}^{d \times d}$, by minimising:

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

- **orthogonality constraint:**
  — $\mathbf{E}_\tau \approx \mathbf{E}_\sigma \mathbf{W}$ and $\mathbf{E}_\sigma \approx \mathbf{E}_\tau \mathbf{W}^\top$
  — improves the performance of machine translation[2]
- **solution:** $\mathbf{W} = \mathbf{V}\mathbf{U}^\top$, where $\mathbf{E}_\tau^\top \mathbf{E}_\sigma = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ (SVD)

[1] Smith et al. (2016) [2] Artetxe et al. (2016)

Step 2 – Semantic similarity (Θ_s)

Compute a query (source) to query (target) similarity matrix

- source, target query embedding: $\mathbf{e}_{q_i}$, $\mathbf{e}_{q_j} \in \mathbb{R}^{1\times d}$
- cosine similarity matrix $\mathbf{\Omega} \in \mathbb{R}^{s\times|\mathcal{P}_\mathsf{T}|}$, $\omega_{ij} = \dfrac{\left(\mathbf{e}_{q_i}\mathbf{W}\,\mathbf{e}_{q_j}^\top\right)}{\left(\|\mathbf{e}_{q_i}\mathbf{W}\|_2\|\mathbf{e}_{q_j}\|_2\right)}$

Step 2 – Semantic similarity (Θ_s)

Compute a query (source) to query (target) similarity matrix

- source, target query embedding: $\mathbf{e}_{q_i}, \mathbf{e}_{q_j} \in \mathbb{R}^{1\times d}$
- cosine similarity matrix $\mathbf{\Omega} \in \mathbb{R}^{s \times |\mathcal{P}_\mathsf{T}|}$, $\omega_{ij} = \dfrac{\left(\mathbf{e}_{q_i} \mathbf{W}\, \mathbf{e}_{q_j}^\top\right)}{\left(\|\mathbf{e}_{q_i} \mathbf{W}\|_2 \|\mathbf{e}_{q_j}\|_2\right)}$

Inverted softmax

- using ω_{ij} directly for translations can generate **hubs**
  — target query is similar to way too many different source queries
  — reduces performance of machine translation[1]
- instead, given a source query q_i, find a target q_j that maximises

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

[1] Dinu et al. (2014); Smith et al. (2016)

Step 2 – Semantic similarity (Θ_s)

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

- α_j: ensures $P_{j \to i}$ is a probability
- s: number of source queries
- η: learned by maximising the log probability over the alignment dictionary ($\sigma \to \tau$): $\underset{\eta}{\operatorname{argmax}} \sum_{\text{pairs } ij} \ln\left(P_{j \to i}\right)$

Inverted softmax

- probability that a target query translates back to the source query
- hub target query $\implies$ large denominator
- top-k target queries are selected as possible mappings of q_i

Step 2 – Semantic similarity (Θ_s)

Inverted softmax

- probability that a target query translates back to the source query
- hub target query $\implies$ large denominator
- top-k target queries are selected as possible mappings of q_i

Determine the semantic similarity score by

- using these top-k queries (average if $k > 1$)
- and computing

$$\Theta_{\mathsf{s}}(q_i, q_j) = \left(\mathbf{e}_{q_i} \mathbf{W}\, \mathbf{e}_{q_j}^{\top}\right) / \left(\|\mathbf{e}_{q_i} \mathbf{W}\|_2 \|\mathbf{e}_{q_j}\|_2\right)$$

Step 2 – Temporal similarity (Θ_c)

Exploit query relationship in the frequency space:

- important relationship; based on the core **statistical** input information

Step 2 – Temporal similarity (Θ_c)

Exploit query relationship in the frequency space:

- important relationship; based on the core **statistical** input information
- compute pair-wise correlation between the frequency time series of source and target queries
- flu seasons may be **offset** in different locations

Step 2 – Temporal similarity (Θ_c)

Exploit query relationship in the frequency space:

- important relationship; based on the core **statistical** input information
- compute pair-wise correlation between the frequency time series of source and target queries
- flu seasons may be **offset** in different locations
  - ✓ compute all correlations using a shifting window of $\pm\xi$ weeks
  - ✓ optimal window l_{ij} (source query q_i, target query q_j) is independently computed for each target query

$$\Theta_{\mathsf{c}}(q_i, q_j) = \rho\Big(\mathbf{x}_i(t), \mathbf{x}_j(t + l_{ij})\Big)$$

Step 3 – Determining weights for target queries

Previous steps

- source query q_i allocated weight w_i
- source query q_i mapped to a set $\mathcal{T}_i$ of $k \geq 1$ target queries

Step 3 – Determining weights for target queries

Previous steps

- source query q_i allocated weight w_i
- source query q_i mapped to a set $\mathcal{T}_i$ of $k \geq 1$ target queries

Weight transfer

- if $k = 1$, directly assign w_i to the single target query
- if $k > 1$, w_i is distributed across the k identified target queries

Step 3 – Determining weights for target queries

Previous steps

- source query q_i allocated weight w_i
- source query q_i mapped to a set $\mathcal{T}_i$ of $k \geq 1$ target queries

Weight transfer

- if $k = 1$, directly assign w_i to the single target query
- if $k > 1$, w_i is distributed across the k identified target queries

Weighting schemes

- uniform: $w'_j = w_i / k$
- based on Θ_{ij}, $j \in \{2, \ldots, k\}$: $w'_j = \dfrac{w_i \Theta_{ij}}{\sum\limits_{q_j \in \mathcal{T}_i} \Theta_{ij}}$

Experiments – Transfer tasks

Source location: United States (**US**)

Target locations

- France (**FR**): from English to French
- Spain (**ES**): from English to Spanish
- Australia (**AU**): from English to English, different hemisphere, greater temporal difference in flu outbreaks

Experiments – Transfer tasks

Source location: United States (**US**)

Target locations

- France (**FR**): from English to French
- Spain (**ES**): from English to Spanish
- Australia (**AU**): from English to English, different hemisphere, greater temporal difference in flu outbreaks

Why choose locations where syndromic surveillance systems exist?

- more robust evaluation at this preliminary stage

Experiments – Data

Search query frequencies from Google

- retrieved from the Google Correlate endpoint
- z-scored (by default)
- weekly rates
- September 2007 to August 2016 (both inclusive)
- $\#$ queries: $34{,}121$ (US), $29{,}996$ (FR), $15{,}673$ (ES), $8{,}764$ (AU)

Experiments – Data

Search query frequencies from Google

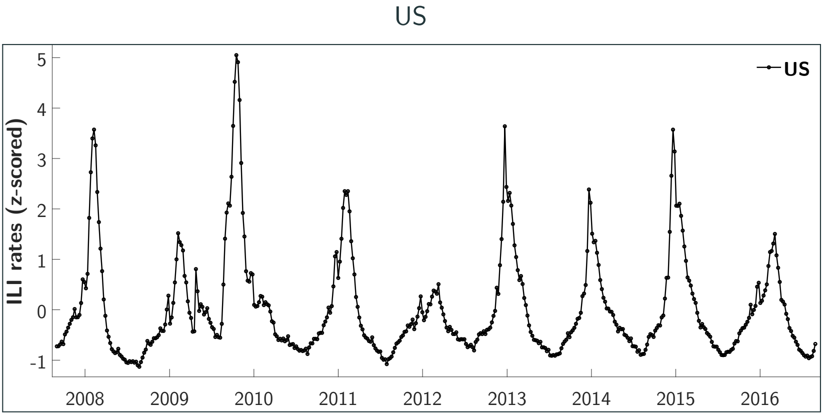
- retrieved from the Google Correlate endpoint
- z-scored (by default)
- weekly rates
- September 2007 to August 2016 (both inclusive)
- # queries: $34{,}121$ (US), $29{,}996$ (FR), $15{,}673$ (ES), $8{,}764$ (AU)

Influenza-like illness (ILI) rates

- data from health organisations in these countries (CDC, SN, SISSS, ASPREN)
- same date range, weekly ILI rates
- z-scored as the metric systems vary in these countries

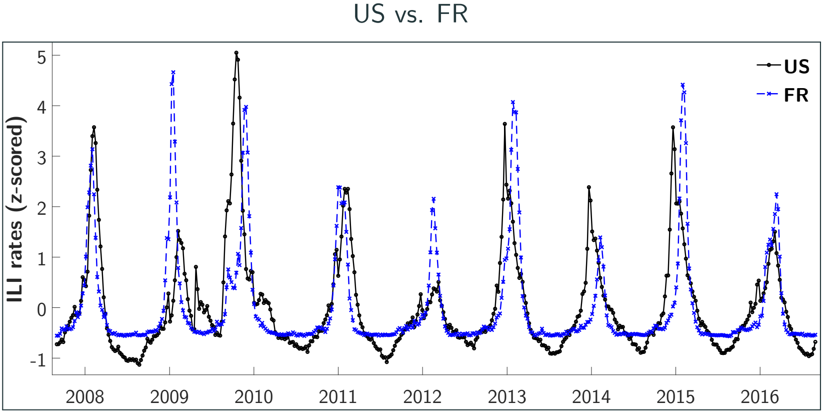
Experiments – ILI rates in the source vs. target country

How similar are they?



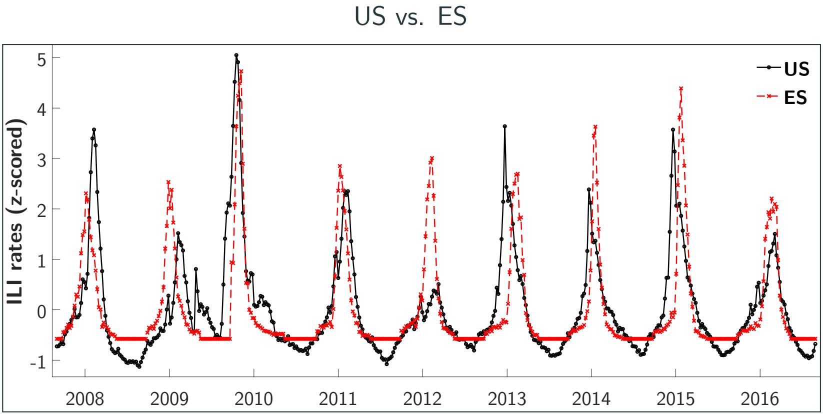
Experiments – ILI rates in the source vs. target country

How similar are they?

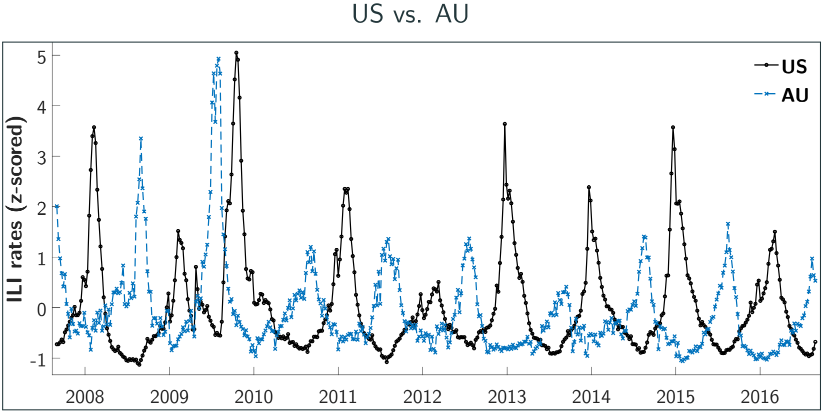


Experiments – ILI rates in the source vs. target country

How similar are they?



How similar are they?



Experiments – Evaluation

Protocol

- train a model using 5 flu seasons, test it on the next
- evaluate performance on the the last 4 flu seasons of our data set
- Θ_{c}: use a window of $\xi = \pm 6$ weeks
- source query $\rightarrow k = \{1, \ldots, 5\}$ target queries
- **Pearson correlation**, mean absolute error (**MAE**), root mean squared error (**RMSE**)

Experiments – Evaluation

Protocol

- train a model using 5 flu seasons, test it on the next
- evaluate performance on the the last 4 flu seasons of our data set
- Θ_c: use a window of $\xi = \pm 6$ weeks
- source query $\rightarrow k = \{1, \ldots, 5\}$ target queries
- **Pearson correlation**, mean absolute error (**MAE**), root mean squared error (**RMSE**)

Baseline models

- **worst case baseline** (**R**): random shuffling of identified query pairs
- unsupervised learning (**U**) using most semantically relevant queries
- **best case threshold** (**S**): supervised learning using elastic net
- transfer component analysis (TCA)[1]

[1] Pan et al. (2009)

Experiments – General observations

In general:

- semantic similarity (Θ_s) is performing better than temporal similarity (Θ_c) when used in isolation
- using semantic or temporal similarity in isolation provides inferior performance, i.e. **hybrid similarity works best**
- values for $k > 1$ **did not help** the hybrid similarity to improve
- when $k > 1$, the non-uniform way of weighting was performing better

Experiments – General observations

In general:

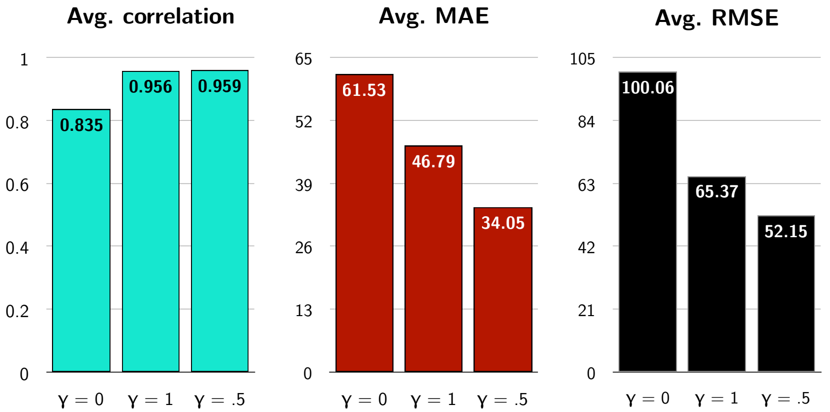
- semantic similarity (Θ_s) is performing better than temporal similarity (Θ_c) when used in isolation
- using semantic or temporal similarity in isolation provides inferior performance, i.e. **hybrid similarity works best**
- values for $k > 1$ **did not help** the hybrid similarity to improve
- when $k > 1$, the non-uniform way of weighting was performing better

Closer look at **results** for $\gamma = 0$, $\gamma = 1$ and the best choice of γ

$$\boxed{\Theta = \gamma\Theta_s + (1-\gamma)\Theta_c,\ \gamma \in [0, 1]}$$

Experiments – Results for France

$$\boxed{\Theta = \gamma\Theta_s + (1-\gamma)\Theta_c,\ \gamma \in [0, 1]}$$

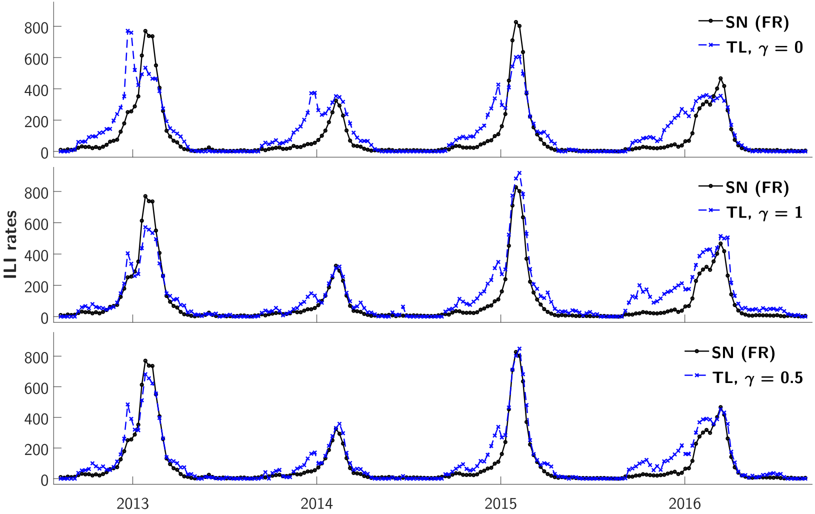


R: 0.911
U: 0.916
S: 0.984

R: 87.729
U: NA
S: 25.088

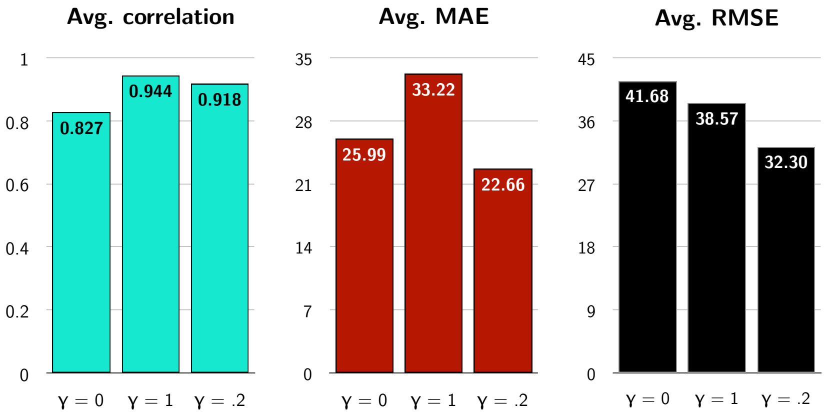
R: 101.845
U: NA
S: 42.349

Experiments – Results for France



Experiments – Results for Spain

$$\boxed{\Theta = \gamma\Theta_{\mathsf{s}} + (1-\gamma)\Theta_{\mathsf{c}},\ \gamma \in [0,1]}$$

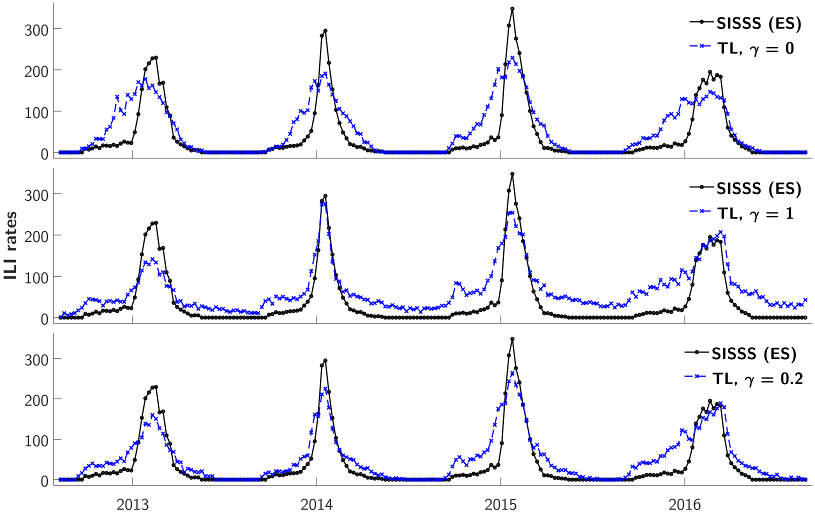


R: 0.872
U: 0.925
S: 0.971

R: 40.311
U: NA
S: 22.120

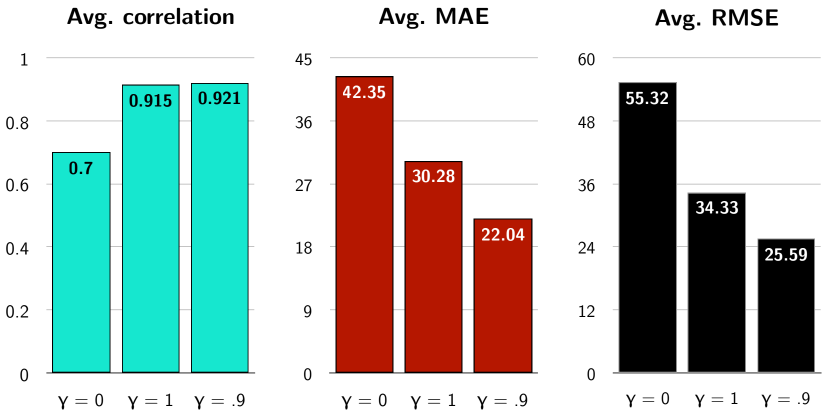
R: 47.204
U: NA
S: 30.600

Experiments – Results for Spain



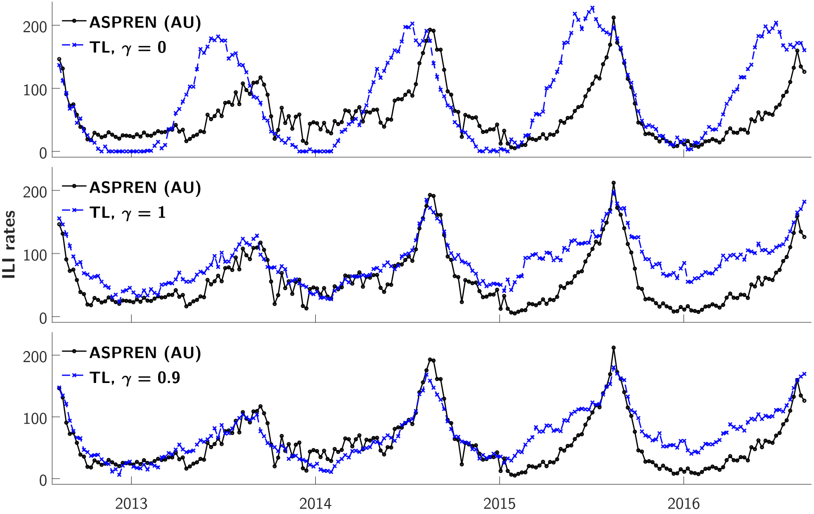
Experiments – Results for Australia

$$\boxed{\Theta = \gamma\Theta_s + (1-\gamma)\Theta_c,\ \gamma \in [0,1]}$$

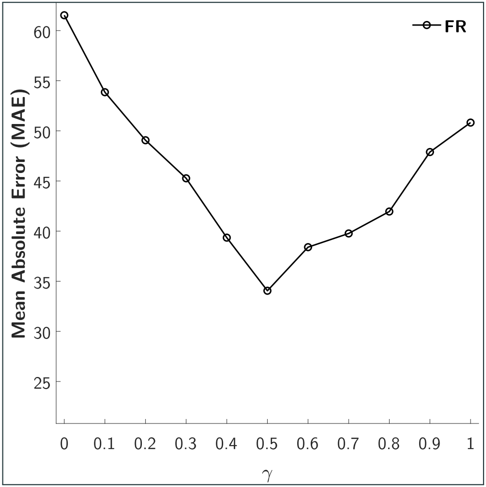


Avg. correlation	Avg. MAE	Avg. RMSE
R: 0.875	R: 25.792	R: 30.080
U: 0.862	U: NA	U: NA
S: 0.916	S: 17.829	S: 21.782

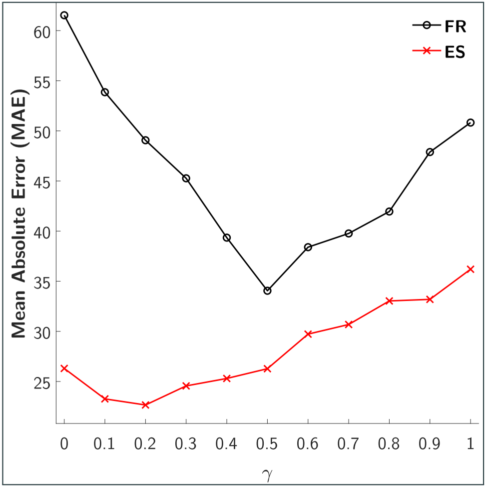
Experiments – Results for Australia



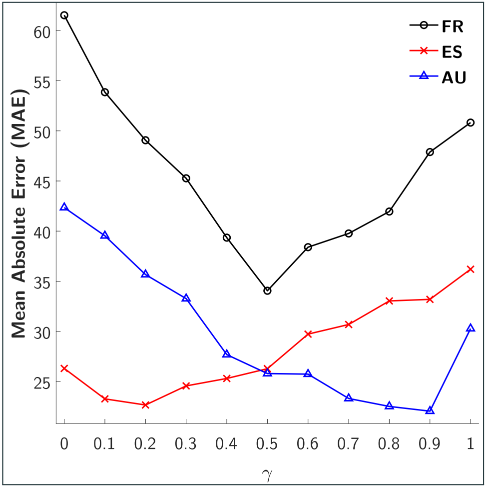
Experiments – Results for different values of γ



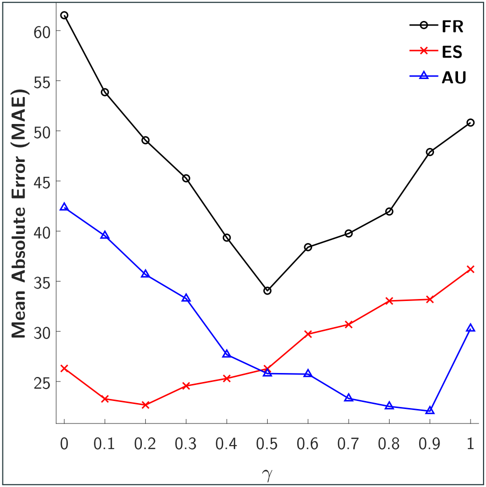
Experiments – Results for different values of γ



Experiments – Results for different values of γ



Experiments – Results for different values of γ



- hybrid similarity optima **differ** per target country
- optimal γ depends on the characteristics of the **input space**
- $\mu(\Theta_c)/\mu(\Theta_s)$ across queries relates to optimal γ: 1.143 (FR), 0.982 (ES), 2.261 (AU)
- identifying optimal γ automatically is an **open task**
- $\gamma = 0.5$ provides better results than non hybrid similarities

Experiments – Where do some of the errors come from?

Error analysis setup

- investigate the models for the optimal gammas
- compute the mean ILI estimate impact (%) during the 10 weeks with highest MAE across all test periods per target country
- identify the worst-5 query pairings

Experiments – Where do some of the errors come from?

Error analysis setup

- investigate the models for the optimal gammas
- compute the mean ILI estimate impact ($\%$) during the 10 weeks with highest MAE across all test periods per target country
- identify the worst-5 query pairings

— – — – — – —

France – *from English (US) to French*

- 24 hour flu $\rightarrow$ **grippe intestinale** (13.24%)
- influenza a treatment $\rightarrow$ grippe traitement (8.07%)
- remedies for colds $\rightarrow$ **rhume de cerveau** (6.75%)
- child temperature $\rightarrow$ **température du corps** (6.37%)
- child fever $\rightarrow$ **fièvre adulte** (6.04%)

Experiments – Where do some of the errors come from?

Error analysis setup

- investigate the models for the optimal gammas
- compute the mean ILI estimate impact ($\%$) during the 10 weeks with highest MAE across all test periods per target country
- identify the worst-5 query pairings

— – — – — – —

Spain – *from English (US) to Spanish*

- mucinez for kids $\rightarrow$ tratmiento de la grippe (20.76%)
- child fever $\rightarrow$ **sinusitis** (7.76%)
- influenza a treatment $\rightarrow$ con gripe (7.02%)
- symptoms pneumonia $\rightarrow$ **bronquitis** (6.04%)
- child temperature $\rightarrow$ temperatura corporal (5.62%)

Experiments – Where do some of the errors come from?

Error analysis setup

- investigate the models for the optimal gammas
- compute the mean ILI estimate impact (%) during the 10 weeks with highest MAE across all test periods per target country
- identify the worst-5 query pairings

— – — – — – —

Australia – *from English (US) to English (AU)*

- 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%)

Conclusions and future work

Summary of outcomes

- **previous efforts** were heavily based on **supervised learning** models
- transfer learning method to enable modelling in areas that lack an established syndromic surveillance system
  — **unsupervised** (no ground truth data at the target location)
  — core operation: how to **map source to target queries**
- **satisfactory performance** (e.g. $r > .92$)
- 21.6% *increase in RMSE* compared to a fully supervised model

Conclusions and future work

Summary of outcomes

- **previous efforts** were heavily based on **supervised learning** models
- transfer learning method to enable modelling in areas that lack an established syndromic surveillance system
  — **unsupervised** (no ground truth data at the target location)
  — core operation: how to **map source to target queries**
- **satisfactory performance** (e.g. $r > .92$)
- 21.6% *increase in RMSE* compared to a fully supervised model

Future work

- study where target location is a **low or middle income country**
  — harder to evaluate; qualitative analysis by experts
- investigate parameters γ (similarity balance) and k (number of target queries in a mapping) further and **learn them from the data**

Questions

?

Acknowledgements

- Funded by the EPSRC project "i-sense" (EP/K031953/1, EP/R00529X/1)
- SISSS and Amparo Larrauri (Spain) for providing syndromic surveillance data
- Simon Moura and Peter Hayes for offering constructive feedback

References

Artetxe, M., Labaka, G., and Agirre, E. (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*, pages 2289–2294.

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

Cook, S., Conrad, C., Fowlkes, A. L., and Mohebbi, M. H. (2011). Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLOS ONE*, 6(8).

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

Eysenbach, G. (2006). Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *Proc. of AMIA Annual Symposium*, pages 244–248.

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

Lampos, V., Miller, A. C., Crossan, S., and Stefansen, C. (2015a). Advances in Nowcasting Influenza-like Illness Rates using Search Query Logs. *Scientific Reports*, 5(12760).

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

References

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

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., and Simonsen, L. (2013). Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLOS Computational Biology*, 9(10).

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2009). Domain Adaptation via Transfer Component Analysis. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1187–1192.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2016). Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. *arXiv preprint arXiv:1702.03859*.

Wagner, M., Lampos, V., Cox, I. J., and Pebody, R. (2018). The added value of online user-generated content in traditional methods for influenza surveillance. *Scientific reports*, 8(1):13963.

References

Yang, S., Santillana, M., and Kou, S. C. (2015). Accurate Estimation of Influenza Epidemics using Google Search Data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478.

Zou, B., Lampos, V., and Cox, I. J. (2018). Multi-Task Learning Improves Disease Models from Web Search. In *Proceedings of the 2018 World Wide Web Conference*, pages 87–96.

Zou, B., Lampos, V., Gorton, R., and Cox, I. J. (2016). On Infectious Intestinal Disease Surveillance using Social Media Content. In *Proceedings of the 6th International Conference on Digital Health*, pages 157–161.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.