Learning Temporal Political Entity Embeddings That Capture Dyadic International Relations from News Data

Qi Pan and **Brendan O'Connor**
College of Information and Computer Sciences
University of Massachusetts Amherst, MA
{qipan,brenocon}@cs.umass.edu

Abstract

Online news provides a rich source of text about international relations and political entities, but who has the time to read it all? We formulate the Dyadic Entity Predicate (DEP) model to learn the distribution of actions between political entities across time from a large corpra of news data in an unsupervised way. The DEP model also learns embeddings that shift smoothly over time for each political entity representing their behavior as the source or receiver of an action. We quantitatively validate the quality of these embeddings by showing that they perform better as latent features for country to country trade level prediction and conflict detection than ground truth non-latent features. To qualitatively assess our model, we build a novel interactive visualization tool to show that the DEP model's output distribution of actions semantically aligns with the opinions of international relations scholars on a number of case studies.

Replication software and data is available at: https://www.github.com/qipanda

1 Introduction

Gone are the days when news was disseminated through physical paper alone. Ubiquitous access to the internet and electronic devices in developed nations have provided a fast and convenient medium for their citizens to absorb the news (Ahlers, 2006). Online news provides the opportunity to computationally study the massive digital textual data they leave behind and the insights they reveal about events in our world. We present an unsupervised way of learning mathematical embeddings (real valued vectors) of countries and international organizations (collectively referred to as political entities) that evolve over time. These are learned in the context of capturing directed dyadic actions between political entities.

In our experiments, we quantitatively show that the embeddings our Dyadic Entity Predicate (DEP) model learns can be used as a set of latent features for tasks such as dyadic trade-level prediction and conflict detection while performing better than when using ground truth descriptive features like GDP. We also show that the DEP model's output allows us to quantify and visualize the shift in how different political entities act towards other entities across time periods without having to manually read and digest millions of news articles. Although convenient to learn, the learned embeddings and their applications are not meant to supplant traditional political science analysis of international relations. We instead intend for the embeddings to support analysis on an empirical level and also enable non-expert users to better understand international relations without having to do a manual in-depth reading of news text on entities they are unfamiliar with.

The goals and data we use build on the prior work of (O'Connor et al., 2013) which uses the same English news data between 1987-2008. Our modelling work is primarily inspired by the Skip-gram model (Mikolov et al., 2013) and work exploring the properties/training of word embeddings over time (Bamler and Mandt, 2017; Hamilton et al., 2016; Yogatama et al., 2011). Details regarding the news data are described in section 2. We formally introduce our DEP model in section 3 which combines two political entity embeddings in a dyadic relationship in a novel way and uses predicate based dependency paths as context instead of the typical sliding window bag of words for Skip-gram models. In section 4 we describe our training procedure for our DEP model and in section 5 we show the results of our experiments/evaluation of our model. Finally, we talk about related works in section 6.

2 Data

Our model, the DEP model is described in section 3 and is trained on 6.5 million news articles from the English Gigaword 4th edition corpus (Parker et al., 2009). This is supplemented by a small sample of articles from the *New York Times*, Annotated Corpus (Sandhaus, 2008). This is the same set of data used in (O'Connor et al., 2013) and the same preprocessing techniques are applied to it to create a workable dataset of 365,623 tuples with the form:

$$\langle s, r, p, t \rangle$$

Where s and r are "source" and "receiver" political entities (mainly countries), p is a predicate path which is the shortest dependency path between s and r in a sentence containing the predicate (typically a verb), and finally t is a timestep (e.g. calendar date) representing when this tuple was published in the news. These tuples represent who (s) did what (p) to whom (r) and when (t). For example, a excerpt from an article written by the Associated Press Worldstream on April 1st 2002 stated :

> "...**Israeli** forces *entered* biblical Bethlehem and another **West Bank** town, Qalqiliya."

This generated the tuple:

$$\langle \textbf{ISR, PSE,} \; enter \xleftarrow{dobj}, 196 \rangle$$

Where 196 is simply a unique ID for the Year-Month combination of April 2002 (In this case the granularity of t is monthly). For the full details on how these tuples were extracted from the raw corpus, please refer to the Data section of (O'Connor et al., 2013).

2.1 Political Entity, Time, and Predicate Coverage

All source entities, receiver entities, and predicate paths belong to the sets E_s, E_r and P respectively. The size of these sets are shown in table 1. The number of unique timesteps depends on the granularity we choose for the model, with the finest grain available being per day (as restricted by the corpus data). The timerange of our data goes from 01/01/1987 to 31/12/2008 (DD/MM/YYYY format). As an example, if we chose to have monthly granularity then $t \in [1, 264]$ for a total of 264 unique timesteps (because there are 264 months in our datasets timerange). Let the maximum integer timestep be T (in this example, $T = 264$). The counts of tuples across our dataset's date range can be seen in fig. 1; the low frequencies before 1994 are because only NYT articles are available then and it's sample size is much smaller relative to the Gigaword corpus (see table 2).

Set	Number of Unique Elements
E_s	67
E_r	78
P	10,419

Table 1: The number of unique source entities, receiver entities, and predicate paths in the dataset.

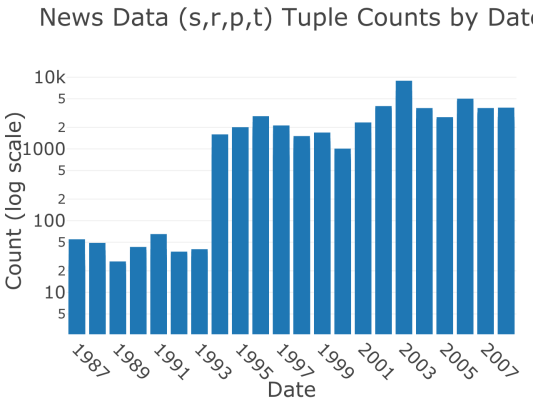


Figure 1: Distribution of tuple counts extracted from our news corpus (Gigaword + NYT) over full date range, counts are given in a log scale.

Corpus	Start Date	End Date	Tuples
Gigaword	1994-01	2008-12	359,502
NYT	1987-01	2007-12	6,121

Table 2: Date ranges and the number of extracted tuples from the two news corpora used

3 Model

In the vanilla Skip-gram model first described in (Mikolov et al., 2013) the distribution of "context" words c for a given "target" word w are modelled to give:

$$\Pr(c|w) = \text{Prob. } c \text{ appears in sliding window of } w \quad \forall c, w \in \mathcal{V} \tag{1}$$

Both c and w come from the same vocabulary, $\mathcal{V}$ and context words in the training data are typically gathered using simple sliding windows around the target word in a given text corpus. All words in $\mathcal{V}$ have an embedding, a K dimensional real valued vector, as both a target and context word. If $W \in \mathbb{R}^{\mathcal{V}\times K}$ is the collection of target word embeddings and $C \in \mathbb{R}^{\mathcal{V}\times K}$ is the collection of context word embeddings then the Skip-gram model output is calculated as:

$$\Pr(c|w) = \frac{\exp(W_{w,\cdot}C^T_{c,\cdot})}{\sum_{c'\in\mathcal{V}}\exp(W_{w,\cdot}C^T_{c',\cdot})} \tag{2}$$

We modify the "context" and "target" in the Skip-gram model to fit our data such that $c \equiv p$ and $w \equiv \langle s, r, t\rangle$. Therefore, the probability we model can be interpreted as:

$$\Pr(p|s,r,t) = \text{Prob. that } s \text{ did } p \text{ to } r \text{ at time } t \quad \forall s \in E_s, \forall r \in E_r, \forall p \in P, \forall t \in [1,T]\cap\mathbb{Z} \tag{3}$$

It is straight forward to represent our predicates as a collection of K dimensional embeddings, let this collection be $V \in \mathbb{R}^{|P|\times K}$. V plays an analogous role to C in the vanilla Skip-gram model and contains one K dimensional embedding per predicate in our dataset. Creating the analogous representation of W is less straight forward as multiple parameters, s, r, and t, determine each embedding. We assigned each $\langle s, t\rangle, \forall s \in E_s, \forall t \in [1,T]\cap\mathbb{Z}$ and $\langle r, t\rangle, \forall r \in E_r, \forall t \in [1,T]\cap\mathbb{Z}$ its own embedding of size $K/2$ (K is always chosen to be even) and concatenate the two $K/2$ embeddings to create a single K dimensional embedding per $\langle s, r, t\rangle$. Let $Q \in \mathbb{R}^{|S|\times T\times K/2}$ and $U \in \mathbb{R}^{|R|\times T\times K/2}$ be the collection of embeddings for both source and receivers per timestep. Our model output is calculated as:

$$\Pr(p \mid s,r,t) = \frac{\exp(\left[Q_{s,t,\cdot} \quad U_{r,t,\cdot}\right] V^T_{p,\cdot})}{\sum_{p'\in P}\exp(\left[Q_{s,t,\cdot} \quad U_{r,t,\cdot}\right] V^T_{p',\cdot})} \tag{4}$$

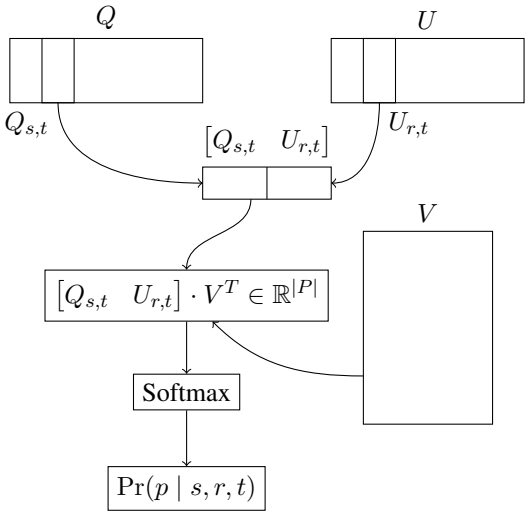


Figure 2: Graphical representation of DEP model for a given $\langle s, r, t\rangle$, the output is a distribution across all predicate paths in P

We call this the Dyadic Entity Predicate (DEP) model because it models the distribution of predicate paths between dyads of entities. A graphical depiction of the model can be seen in fig. 2.

3.1 Assumptions about Int. Relations

Our modelling choices lead to inherent assumptions made about international relations in the real world. These are important to know when interpreting the model, some of the main ones are:

1. The meaning of predicates do not change over the time period of the dataset as V has no time component

2. Political entities act differently depending on if they are the "source" or "receiver" of an interaction since an entity which exists in $E_s \cap E_r$ will have different embeddings for a given t in Q and U

3. A political entity both acts similarly to all other entities as a source and receives interactions similarly from other entities as a receiver since the corresponding embedding at a given t in Q and U is the same no matter what other embedding is paired it as the model input

Concatenation has the advantage of allowing political entities to share information within it's embeddings across different pairings which helps alleviate the sparsity of specific $\langle s, r\rangle$ pairs. This is one of the main motivations for concatenation

between s and r as opposed to a unique embedding per $\langle s, r\rangle$ pair. For example, the pair $\langle s$=USA, r=CHN, $t\rangle$ is much more common than $\langle s$=USA, r=KSV, $t\rangle$ but through concatenation, the USA source embedding...

4 Learning and Building Representative Embeddings

Similair to how the original goal of the Skip-gram model is to learn good representations of it's word embeddings W and C, the goal of the DEP model is to learn good representations of political entities (Q and U) and the actions they take (V). Instead of explicitly trying to find these representations, we train our model by adjusting its model parameters (the embeddings in Q, U, and V) to maximize the likelihood that we will observe our training data. In other words, if we observe the tuple $\langle s, r, p, t\rangle$, then we want to increase our model's output for $\Pr(p \mid s, r, t)$ to better reflect real word news reports. Cross-entropy loss can be used to match this intention; let N be the number of samples in the training set, $y^{(i)} \in \mathbb{R}^K$ be a one-hot vector where the kth element is one and corresponds with the ith training sample's predicate p. The ith sample is a tuple $\langle s_i, r_i, p_i, t_i\rangle$ where $s_i \in S, r_i \in R, p_i \in P, t_i \in [1, T] \cap \mathbb{Z}$. Finally, let $\theta = \{Q, U, V\}$ be all the model parameters of DEP model; the negative log-likelihood loss becomes:

$$\begin{aligned}\mathcal{L}_{NLL}(\theta) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k'=1}^{K} y_{k'}^{(i)} \log(\Pr(p_{k'} \mid s_i, r_i, t_i)) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \log(\Pr(p_i \mid s_i, r_i, t_i))\end{aligned} \tag{5}$$

Minimizing eq. (5) effectively maximizes the mean log-likelihood of the observed predicates across all training tuples of the form $\langle s, r, t\rangle$. This is both standard for Skip-gram models in (Mikolov et al., 2013) and in multiclass supervised learning problems such as image classification in the pioneering paper (Krizhevsky et al., 2012).

4.1 Political Entity Embeddings Across Time

A desirable property of Q and U is that for a given entity $e \in E_s \cup E_r$ it should have a similar embedding at time t and $t+1$ as either a source (Q) or receiver (U). This is to match our intuition that the poltical entity e carries over a portion of it's behavior from the previous time step and changes it's behaviour in a smooth fashion between time steps. If we only optimized with respect to eq. (5) this property is not accounted for and we may not see smooth transitions or any continuity between an entities source or receiver embeddings over time. To try to obtain this property, we L2 regularize on the one time step difference for each $e \in E_s \cap E_r$ in the vein of (Yogatama et al., 2011) as shown in eq. (6). This will penalize the model for reacting to harshly to highly variant data points between time steps if there isn't enough evidence to justify doing so.

$$\begin{aligned}\mathcal{L}_{smooth}(Q, U) = \alpha \sum_{t=2}^{T} \Bigg(&\sum_{s \in S} \|Q_{s,t} - Q_{s,t-1}\|_2^2 + \\ &\sum_{r \in R} \|U_{r,t} - U_{r,t-1}\|_2^2 \Bigg)\end{aligned} \tag{6}$$

The smoothing hyperparameter α determines the smoothing strength, or how resistant we want to be to contrary data one time step ahead. The full loss function becomes eq. (7).

$$\mathcal{L}_{total}(\theta) = \mathcal{L}_{NLL}(\theta) + \mathcal{L}_{smooth}(Q, U) \tag{7}$$

4.2 Optimization

Optimization is done with mini-batch gradient descent using a step size of 1, a batch size of 32, and over 50 epochs where the batches are shuffled every epoch. L2 regularization in eq. (6) was chosen partially because the gradient is easily calculated for gradient descent. For the experiments and evaluations in section 5 the DEP model with five smoothness levels (α) were trained on both annual and monthly granularity and $K = 300$, giving a total of ten trained models. The respective loss curves for eq. (7) are shown in fig. 3. Visually, each model seems to have converged to some local minima over 50 epochs.

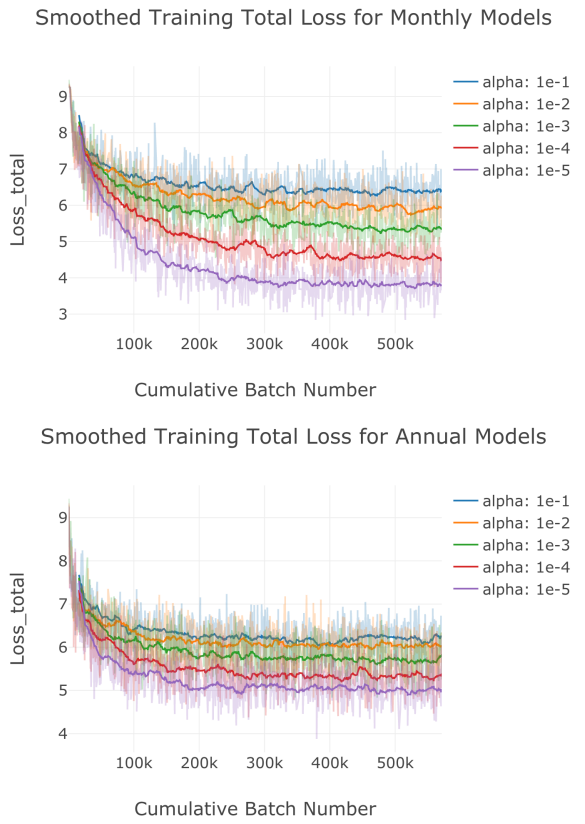


Figure 3: Smoothed training loss curves for monthly (above) and annual (below) granularity over 50 epochs

5 Evaluation

To verify the quality our model's political entity embeddings $\{Q, U\}$ with respect to accurately representing international relations, we show quantitative evidence that they are good features for predicting dyadic trade levels and detecting military conflict. We also qualitatively show that the most probable outputs of the DEP model correspond with real world events.

5.1 Predicting Dyadic Trade

Do our political entity embeddings Q and U capture information about international country to country trade? One way to evaluate this is to use Q and U as features for predicting trade levels between countries. In (Westveld et al., 2011), international trade prediction was done on a dataset consisting of bilateral trade values for 58 different countries between the years 1981-2000. Each row has a tuple of features $x = \langle e, i, t, G_e, G_i, D_{e,i}, Pol_e, Pol_i, CC_{e,i} \rangle$ representing the exporting country (e), importing country (i), year (t), log GDP of country e and i (G_e and G_i), the physical distance between country e and i ($D_{e,i}$), Polity of country e and i (Pol_e and Pol_i), the Cooperation in conflict score between country e and i ($CC_{e,i}$). Each tuple of data also has an associated log trade value of $y \in \mathbb{R}$. Polity is a measure of how democratic a country is on an integer scale between $[0, 20]$ (Marshall et al., 2017) and CC is a categorical measure of whether the pair cooperated in a dispute ($+1$), did nothing together that year (0) or were on opposite sides of a dispute (-1). To match corresponding features from Q and U to each of these tuples, we treat sources as exporters ($s \equiv e$), receivers as importers ($r \equiv i$), and use a year granularity version of our DEP model. The dataset for our trade-prediction experiments consists of all tuples in the trade data where both $Q_{s,t}$ and $U_{r,t}$ exist our model, this is summarized in **??** (12.20% of the original trade data overlaps with features from our DEP model). Let this dataset be $X_{inter} \in \mathbb{R}^{N_{trade} \times 9}$ where each row consists of the 9 features of x and let $Y_{inter} \in \mathbb{R}^{N_{trade}}$ be the corresponding log trade values.

Using the methodology from (Westveld et al., 2011), we randomly split the intersecting dataset into a 75-25 train-test split. Let I_{train} and I_{test} be the sets of training and testing tuple indices respectively. Using Q and U, we train a simple linear regression model on $X_{inter}[I_{train}]$ and $Y_{inter}[I_{train}]$ to predict $y \in Y_{inter}[I_{test}]$. Our prediction given a tuple of test features $x \in X_{inter}[I_{test}]$ is shown in eq. (8) as $\hat{y}$ where $\beta \in \mathbb{R}^{K+1}$ are the coefficients per feature and the bias term.

DEP model:

$$\hat{y}(x) = \beta \cdot \begin{bmatrix} 1 & Q_{i,t} & U_{e,t} \end{bmatrix} \quad (8)$$

We compare our linear model with the linear model used in (Westveld et al., 2011) shown in eq. (9) and a baseline of just predicting the mean y of the training data as defined by eq. (10). The Westveld-Hoff model is built upon typical "gravity" models for international trade which uses the idea that the trade level between two countries is proportional to the size of their economies (e.g. GDP) (Tinbergen, 1962).

Westveld-Hoff Model:

$$\begin{aligned} \hat{y}_{\text{hoff}}(x) = \beta_0 &+ \beta_1 G_e + \beta_2 G_i + \beta_3 D_{e,i} + \\ &\beta_4 Pol_e + \beta_5 Pol_i + \\ &\beta_6 CC_{e,i} + \beta_7 \left(Pol_e \times Pol_i\right) \end{aligned} \quad (9)$$

	Our Data	Trade Data	Intersection w/ DEP model
unique s/e	67	58	26
unique r/i	78	58	23
unique years	22	20	14
year range	1987-2008	1981-2000	1987-2000
tuples	365,623	66,120	N_{trade} = 8,064 (12.20% of original trade data)

Table 3: Intersection of the original trade data with what our model can provide features for (and therefore make predictions on)

Baseline Mean Model:

$$\hat{y}_{\text{mean}}(x) = \frac{\sum_{j \in I_{train}} Y_{inter}[j]}{|I_{train}|} \quad (10)$$

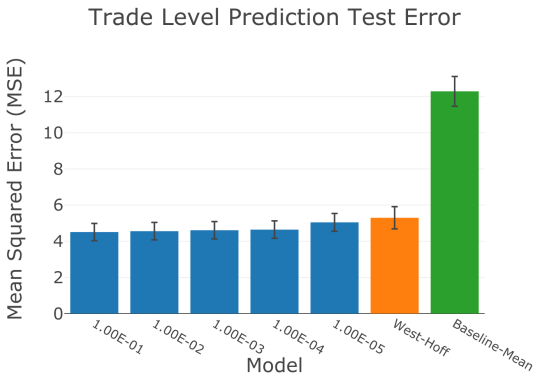
We use mean squared error (MSE) on the test set to evaluate all models. To alleviate variance from the randomized 75-25 split, we run the experiment on 1000 independent and randomized 75-25 splits and aggregate the results. The results can be seen in fig. 4; overall our models across all smoothness levels (α) do slightly better than the Westveld-Hoff model with the baseline mean prediction model doing much worse. This provides evidence that our country entity embeddings $\{Q, U\}$ are at least as good or better than the features in x (e.g. $Poly$ and D) for capturing information about trade levels between countries. The more remarkable result is that only 644 of the 8,064 ($\approx 8\%$) tuples in X_{inter} have a corresponding $\langle e \equiv s, e \equiv r, t\rangle$ which exists in our news corpus data, the rest of the provided features from our model embeddings $\{Q, U\}$ are interpolations of the DEP model.

5.2 Detecting Dyadic Military Conflict

To what extent do our DEP model embeddings $\{Q, U\}$ capture a political entities tendency to be engaged in a militarized dispute with another entity? To evaluate this, we refer to records of such disputes between pairs of countries recorded in the Dyadic Militerized Interstate Dispute (Dyadic MIDs) dataset (Maoz et al., 2019). The Dyadic MIDs dataset is one of the most prominent datasets used to study international relations and has been used to study whether economic ties and formal alliances facilitate interstate peace or conflict (Leeds, 2003; Barbieri, 1996). In our experiment we will see if $\{Q, U\}$ are good features for building a classifier that is able to detect conflict in the Dyadic MIDs dataset.

Each entry in the Dyadic MIDs dataset represents a pair of countries involved in a specific conflict. Each row contains many features but for our experiments it is just important to note the features $\langle a, b, tr, h, role_a, role_b\rangle$ representing the countries a and b, the date time range of the conflict tr, the hostility level h, and the roles of the countries in this conflict $role_a$ and $role_b$. The target variable is $h \in \{1, 2, 3, 4, 5\}$ which we convert to a binary label $y = \mathbb{1}[h \geq 4]$ in the same fashion as (O'Connor et al., 2013). For each tuple $\langle a, b, tr, y, role_a, role_b\rangle$, we explode it into copies of itself replacing the time timerange tr with each unique month t in tr. For example, if tr was Jan 3rd 1996 to Feb 20th 1997, there would be 14 copies of the tuple made each with a unique month $t \in \{\text{Jan-1996}, \text{Feb-1996}, \ldots, \text{Jan-1997}, \text{Feb-1997}\}$. The last thing to do so that we can match our features $Q_{s,t}$ and $U_{r,t}$ to the Dyadic MIDs data is to match s with a and r with b. To ensure that we preserve the semantics of s being a "source" and r being a "receiver" of political action, we filter down the Dyadic MIDs dataset such that a is either the initiator of the conflict or joined on the initiators side and b is the target or joined the target's side (technically this means $role_a \in \{\text{Primary Initiator}, \text{Joiner on initiator side}\}$ and $role_b \in \{\text{Primary Target}, \text{Joiner on target side}\}$). This gives us a dataset with tuples $\langle s, r, t, y\rangle$ which we can easily align with the our features $\{Q, U\}$. Any missing Dyadic MIDs entries for a specific (s, r) pair within our DEP model's predictive range (Jan 1987 to Dec 2008) are presumed to have $y = 0$ (no conflict). To make sure no (s, r) dyad is trivial to detect, only dyads with at least one conflict and one non-conflict month over the predictive time range are included. This gives us 126 unique dyads across all 256 months in the time range making for $126 \times 256 = 32,256$ tuples of data.

A randomized 80-20 train-test split is performed on the tuples for fitting and testing. Our



Model	MSE mean	MSE std.
DEP $\alpha = 10^{-1}$	**4.5088**	**0.4786**
DEP $\alpha = 10^{-2}$	4.5568	0.4828
DEP $\alpha = 10^{-3}$	4.6087	0.4815
DEP $\alpha = 10^{-4}$	4.6424	0.4829
DEP $\alpha = 10^{-5}$	5.0445	0.4925
Westveld-Hoff	5.2977	0.6141
Baseline Mean	12.2880	0.8216

Figure 4: Mean test MSE's for international country to country trade prediction of various models over 1000 randomized 75-25 train-test splits. Error bars show one standard deviation across the 1000 splits.

baseline model is to predict the mean y value for the input (s, r) in our training data. Let $X_{train} \in \mathbb{R}^{N_{train} \times 3}$ be the $\langle s, r, t \rangle$ tuples in our training set and $Y_{train} \in \mathbb{R}^{N_{train}}$ be the corresponding binary hostility labels, the baseline model's prediction on the probability of conflict is given by eq. (11).

$$\hat{y}_{baseline}(s, r, t) = \frac{\sum_{t=1}^{T} Y_{train}^{(s,r,t)}}{T} \tag{11}$$

For our classifiers, we train a logistic regression classifier with L1 regularization on our training data X_{train}, Y_{train} using features from our DEP model with five different smoothness levels α. Our classifiers makes predictions on the probability of conflict according to eq. (12) where $\beta \in \mathbb{R}^{1+K}$ are the learned model parameters. Choosing the L1 regularization strength λ was done with 5-folds cross validation.

$$\hat{y}(s, r, t) = \frac{1}{1 + \exp(-\beta^T \begin{bmatrix} 1 & Q_{s,t} & U_{r,t} \end{bmatrix})} \tag{12}$$

The results on the test set are shown in table 4 using ROC Area Under the Curve to see the diagnostic ability of the binary classifiers across all thresholds. Overall, our best model outperforms the baseline by a decent ROC AUC margin of ≈ 0.085 indicating that the learned embeddings $\{Q, U\}$ from our DEP model contains dyadic hostility information that is more indicative of conflict than the historical mean conflict level between the countries.

5.3 Correspondence with Real World International Relations

For a given dyad, does the distribution of predicate paths semantically make sense over time? To qualitatively explore this, we have built an interactive visualization tool which plots the predicate path distribution $\Pr(p \mid s, r, t)$ our DEP model outputs (the full UI can be seen in fig. 5). We will show how the output distribution in four dyads align semantically with contemporary opinion on specific dyadic international relations. Because there are over 10,000 predicate paths in P, we will only plot the "top" ones for clarity in our visual evaluations. Top predicate paths are those with the highest average $Pr(p \mid s, r, t)$ for a given dyad (s, r) across all times steps $t \in [1, T] \cap \mathbb{Z}$.

Model	Best λ	Test ROC AUC
DEP $\alpha = 10^{-1}$	10^{-1}	0.8602
DEP $\alpha = 10^{-2}$	**10^{-4}**	**0.9017**
DEP $\alpha = 10^{-3}$	10^{-1}	0.8922
DEP $\alpha = 10^{-4}$	10^{-2}	0.8561
DEP $\alpha = 10^{-5}$	10^{-7}	0.7650
Dyad Mean	N/A	0.8148

Table 4: Dyadic MIDs conflict detection test ROC AUC's using a L1 regularized logistic classifier trained with DEP model features across varying smoothness settings α. Performance of the baseline model (predicting mean hostility within a dyad) is also shown.

Case 1: USA→RUS Relations: In our first case study, we plot the top four predicate paths from the dyad (s=USA,r=RUS) using our visualization tool (fig. 6). The probability of predicate paths related to cooperation $p \in$ {"meet with", "visit"} have the highest likelihood early on, but taper off towards the end. This is in contrast with the predicate paths related to distrust and slander $p \in$ {"accuse", "critisize"} which rise at the end. These trends align with analysis done by (Savranskaya, 2018) describing America's warmer rela-

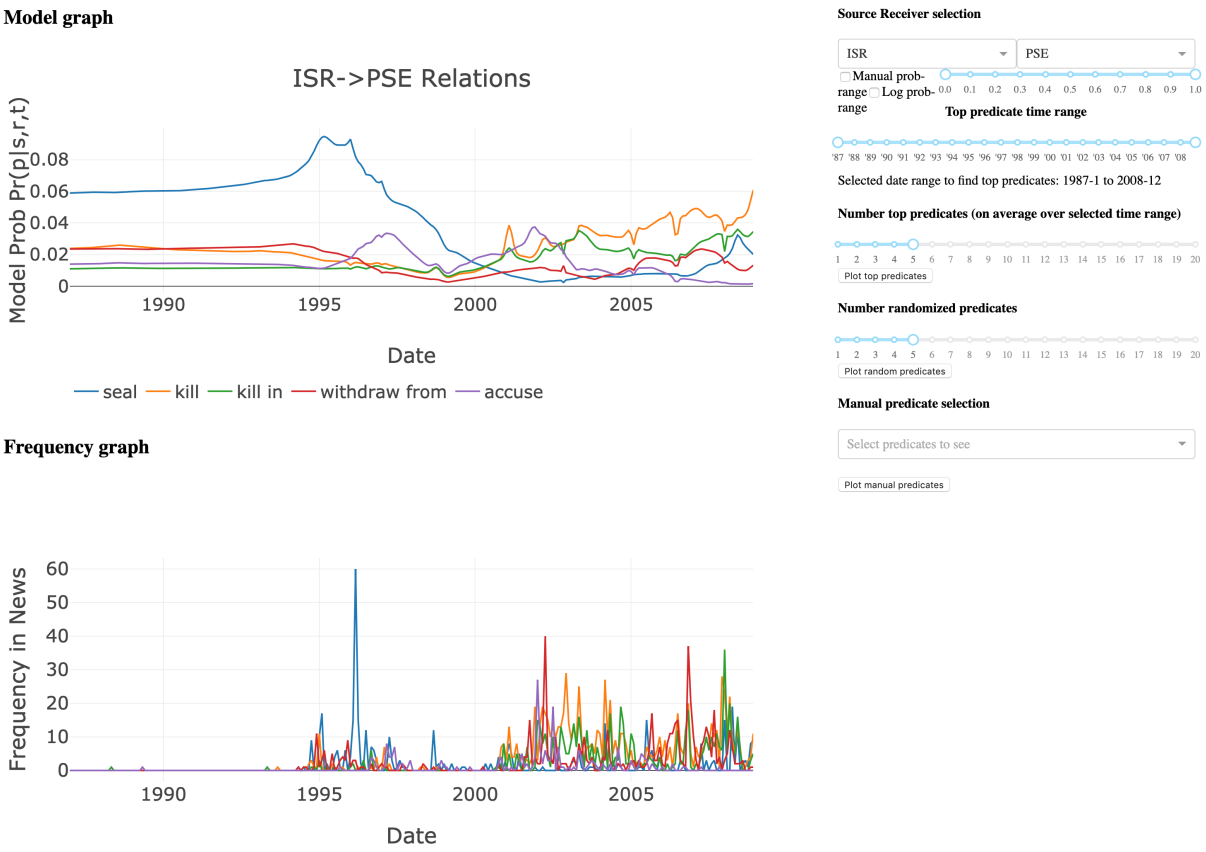


Figure 5: Interactive Visualization Tool for DEP model output, able to adjust the source, receiver, set of predicate paths, and time range to see. Can also get top predicate paths (highest average distribution probability over time range) and randomized predicate paths.

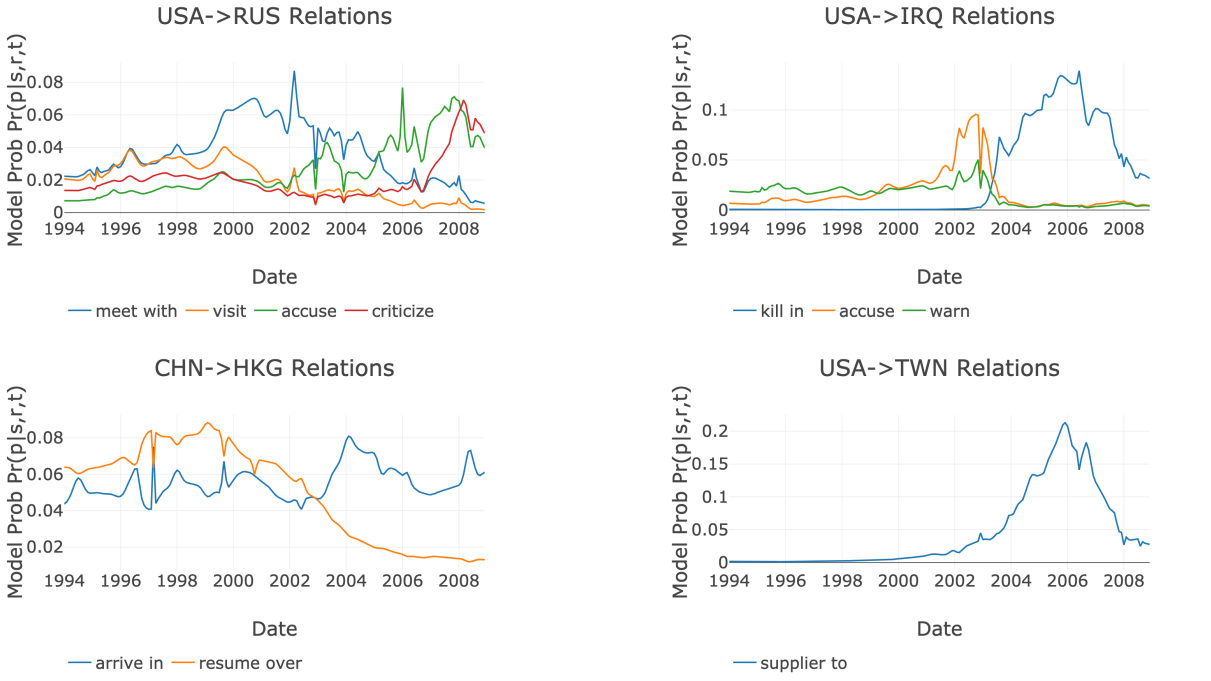


Figure 6: Time series of model probability $\Pr(p \mid s, r, t)$ for the top predicate paths in the USA→RUS (top left), USA→IRQ (top right), CHN→HKG (bottom left), and USA→TWN (bottom right) dyads.

tionship with Russia in the 1990's when relationships between U.S. President Bill Clinton and inaugural Russian President Boris Yeltsin were amicable. Clinton is even noted as having helped Yeltsin be reelected in 1996. U.S. Russia relations however, were notably strained in the 2000's which corresponding with the negative predicate paths rising in the 2000's. The relationship was so poor that when Barack Obama became president he decided to try and officially "reset" negative relations with Russia in 2009 (Stoner and McFaul, 2015).

Case 2: USA→IRQ Relations: Our second example looks at the top three (USA, IRQ) predicate paths shown in fig. 6. Following the 2001 September 11th attacks and leading up to the 2003 American invasion of Iraq, threatening and accusatory predicate paths ("accuse" and "warn") dominate until the invasion occurs where the "kill in" predicate path becomes the most probable. The accusatory trends match with the Bush administrations actions at the time which attempted and succeeded in publicly building a case for the invasion of Iraq predicated on Saddam Hussein developing weapons of mass destruction and being a threat to the world (Byers, 2004). The "kill in" trend which starts alongside the 2003 invasion also makes sense given the massive amount of Iraqi causalities over the course of the conflict with an estimated 401,000 deaths due to the war between 2003-2011 (Hagopian et al., 2013).

Case 3: CHN→HKG Relations: Our third case looks at how the top two predicate paths for the dyad (CHN, HKG) change. Hong Kong became a colony of Great Britain as a result of the First Opium Wars in 1842 but was officially handed back to the People's Republic of China in 1997 (Carroll, 2007). The "resume over" predicate path (indicating the prevalence of reports like "China resumes control over Hong Kong") also peaks around 1997 which semantically aligns with the hand over of Hong Kong back to mainland China. As noted by (Law and Lee, 2006), many Hong Kongnese also have children who live in mainland and so the immigration of mainlanders into Hong Kong is a regular contentious topic. This matches the second place ranking of the "arrive in" predicate path (e.g. "Chinese arrive in Hong Kong") throughout the whole time range.

Case 4: USA→TWN Relations: Finally, our fourth case looks the top predicate path, "supplier to", for the (USA, TWN) dyad. This corresponds to significant arms sales from the U.S. to Taiwan throughout George W. Bush's presidency from 2000-2008. During this time period, submarines, aircraft, and destroyers were sold to Taiwan who used an annual defense budget ranging between $9-$18 billion USD to purchase them (Kan, 2009). Although not a rigorous analysis of all possible dyads, these case studies indicate that the model distribution of predicate paths do align with many studied real life dyadic international relations between political entities.

6 Related Work

Prior work on computationally discovering political entity attributes includes prediction of country regime types (Minhas et al., 2015) and gathering public sentiment on countries from tweets (Chambers et al., 2015). Research sharing our goal of uncovering dyadic entity to entity international relations in an unsupervised way is studied in (O'Connor et al., 2013) using predicate paths as context and more recently in (Han et al., 2019) using the Relationship Modeling Network (RMN) proposed in (Iyyer et al., 2016) (RMN also learns embeddings for dyadic relationships like our DEP model). There are also more general lines of work which attempt to extract relationships between entities in an unsupervised way instead of predefining them such as topic modelling (Blei et al., 2003; Schein et al., 2015) and template/frame learning (Chambers and Jurafsky, 2011; Cheung et al., 2013).

Our DEP model primarily builds on the idea of word embeddings which evolve based on the context that surronds them using things like the Skip-gram model (Mikolov et al., 2013). Joint optimization of the embeddings across time steps is formalized and explored for word embeddings in (Bamler and Mandt, 2017) and the statistical laws of such embeddings are explored in (Hamilton et al., 2016). Our L2 regularization formulation however, is closest to work in (Yogatama et al., 2011) which talks about a more general setting than just word embeddings. Using this formulation in our work allows us to largely ignore the pains of undefined gradients when learning and is easily integrated with gradient descent methods.

To evaluate the quality of our learned embed-

dings, we used them as latent features to perform dyadic country level trade level predication and conflict detection. Work on these two tasks generally build models using non-latent real world features for trade level prediction (Westveld et al., 2011; Tinbergen, 1962) and conflict/coup detection (O'brien, 2010; O'Kane, 1981). We show that using our learned embeddings as latent features instead of real world features can perform better for these tasks. We also explored our model outputs with visualization/interactive tools to see if the trends aligned with our knowledge of specific international relations. A similar tool for evaluating large text corpra have been built with a focus on uncovering overall themes and narratives for reporters and users (Handler and O'Connor, 2017).

7 Conclusion

As we move towards a more interconnected and digital world, online news will undoubtedly become a rich source of data for understanding complex international relations. We formulate an unsupervised way of training the DEP model to learn political entity embeddings that shift over time in a smooth manner under the context of capturing directed dyadic actions between political entities reported by the news. Using our DEP model along with dependency path preprocessing, we combine entity embeddings in a novel way to represent the distribution of directed dyadic actions between entities. We quantitatively show that using these embeddings as latent features for trade prediction and conflict detection work better than models which use real ground truth features. We also qualitatively show that the distribution of predicate paths semantically align with academic studies of international relations between dyads using a custom built interactive visualization tool.

Our work indicates that using predicate path preprocessing to get more detailed contexts (in contrast with bag of words style sliding windows) between political entities can lead to useful entity embeddings and more detailed interpretability (a predicate path tells you much more between s and r than a single word). Future work could explore variations/extensions of our DEP model which only scratches the surface of possible formulations using such predicate paths as context. Our work also suggests that the concatenation of entity embeddings (essentially, treating their features independently in the model) in the dyadic setting can produces good results but there are still many other combination methods to explore (e.g. non-linearities). Finally, our method for enforcing embedding smoothness across time only looks one timestep back, more intricate ways to enforce smoothness (e.g. exponential decay) could be explored and evaluated for embedding quality.

Acknowledgements

We'd like to thank Katherine A. Keith, Su Lin Blodgett, and Xiaohan Ding for their helpful discussion and feedback. We'd also like to thank Abram Handler for their help around the implementation of our DEP model.

References

Douglas Ahlers. 2006. News consumption and the new electronic media. *Harvard International Journal of Press/Politics*, 11(1):29–52.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org.

Katherine Barbieri. 1996. Economic interdependence: A path to peace or a source of interstate conflict? *Journal of Peace Research*, 33(1):29–49.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Michael Byers. 2004. Agreeing to disagree: Security council resolution 1441 and intentional ambiguity. *Global Governance*, 10(2):165–186.

John M Carroll. 2007. *A concise history of Hong Kong*. Rowman & Littlefield Publishers.

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 976–986. Association for Computational Linguistics.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. *arXiv preprint arXiv:1302.4813*.

Amy Hagopian, Abraham D Flaxman, Tim K Takaro, Sahar A Esa Al Shatari, Julie Rajaratnam, Stan Becker, Alison Levin-Rector, Lindsay Galway, Berq J Hadi Al-Yasseri, William M Weiss, et al. 2013. Mortality in iraq associated with the 2003–2011 war and occupation: findings from a national cluster sample survey by the university collaborative iraq mortality study. *PLoS medicine*, 10(10):e1001533.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Xiaochuang Han, Eunsol Choi, and Chenhao Tan. 2019. No permanent friends or enemies: Tracking relationships between nations from news. *arXiv preprint arXiv:1904.08950*.

Abram Handler and Brendan O'Connor. 2017. Rookie: A unique approach for exploring news archives. *arXiv preprint arXiv:1708.01944*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Shirley A Kan. 2009. *Taiwan: major US arms sales since 1990*. DIANE Publishing.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kam-yee Law and Kim-ming Lee. 2006. Citizenship, economy and social exclusion of mainland chinese immigrants in hong kong. *Journal of Contemporary Asia*, 36(2):217–242.

Brett Ashley Leeds. 2003. Do alliances deter aggression? the influence of military alliances on the initiation of militarized interstate disputes. *American Journal of Political Science*, 47(3):427–439.

Zeev Maoz, Paul L Johnson, Jasper Kaplan, Fiona Ogunkoya, and Aaron P Shreve. 2019. The dyadic militarized interstate disputes (mids) dataset version 3.0: Logic, characteristics, and comparisons to alternative datasets. *Journal of Conflict Resolution*, 63(3):811–835.

Monty G Marshall, Ted Robert Gurr, and Keith Jaggers. 2017. Polity iv project: Political regime characteristics and transitions, 1800-2017. *Dataset Users' Manual*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Shahryar Minhas, Jay Ulfelder, and Michael D Ward. 2015. Mining texts to efficiently generate global data on political regime types. *Research & Politics*, 2(3):2053168015589217.

Sean P O'brien. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International studies review*, 12(1):87–104.

Rosemary HT O'Kane. 1981. A probabilistic approach to the causes of coups d'etat. *British Journal of Political Science*, 11(3):287–308.

Brendan O'Connor, Brandon M Stewart, and Noah A Smith. 2013. Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1094–1104.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword fourth edition. *Linguistic Data Consortium.* LDC2009T13.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium*. LDC2008T19.

Svetlana Savranskaya. 2018. Yeltsin and clinton. *Diplomatic History*.

Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. 2015. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054. ACM.

Kathryn Stoner and Michael McFaul. 2015. Who lost russia (this time)? vladimir putin. *The Washington Quarterly*, 38(2):167–187.

Jan J Tinbergen. 1962. Shaping the world economy; suggestions for an international economic policy.

Anton H Westveld, Peter D Hoff, et al. 2011. A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, 5(2A):843–872.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 594–604. Association for Computational Linguistics.