Where Did the Political News Event Happen? Primary Focus Location Extraction in Different Languages

Maryam Bahojb Imani
Computer Science Department
The University of Texas at Dallas
Richardson, USA
maryam.imani@utdallas.edu

Latifur Khan
Computer Science Department
The University of Texas at Dallas
Richardson, USA
lkhan@utdallas.edu

Bhavani Thuraisingham
Computer Science Department
The University of Texas at Dallas
Richardson, USA
bhavani.thuraisingham@utdallas.edu

***Abstract*—**
Political news reports are populated all over the world in various languages. It has a great value to automatically detect the geolocation from these reports for better understanding of the associated events. Although various open-source and commercial tools exist to identify geolocation, they fail to identify at a granular level such as locality or city and they do not support most languages. Most of the techniques view the problem in terms of Named Entity Recognition (NER) and identify geolocation information at the country level for a given text. In this paper, we consider English, Spanish and Arabic news articles from different publishers. We define *primary focus location* as the actual location where the event occurred amongst other *focus* locations mentioned in the report. Our aim is to extract the *primary focus location* regardless of the language from articles belonging to different news agencies. We propose a mechanism to identify potential sentences containing *focus* locations using NER. After that, we perform sentence embedding over words from different languages, and then employ a supervised classification mechanism to predict the *primary focus location*. We also perform bias correction over the training data using a suitable adaptation mechanism to reduce the sampling bias in training data. Our method trains a classifier using bias-corrected training data from news articles published by an agency in one language, while testing the model on news articles published by another agency in a different language. Our empirical results when compared to baseline approaches show superior performance on real-world English, Spanish and Arabic news articles.

***Index Terms*—Focus Location Extraction; Sentence Embedding; Bias Correction; Political Event News**

I. Introduction

With numerous news reports generated every day, many applications search and organize these daily generated news stories for analysis. These applications include determining crime pattern locations, predicting the place of protests and political unrest, and identifying the geolocation of natural disasters. Such applications can largely benefit from identifying precise geolocation information in a timely manner to provide better support for decision making.

In this paper, we focus on identifying the associated locality information of a news article. Typically, the term *Location* is used in a variety of contexts, but the term *Locality* is used to describe a more precise area [1]. A news article may contain multiple related localities mentioned in them. These are called *Focus Locations*. However, we aim to identify the place of occurrence of an event. We call this locality *Primary Focus Location*.

by Ola' Audu 399 words 19 May 2014 10:03 All Africa AFNWS English May 19, 2014 (Premium Times/All Africa Global Media via COMTEX)
-- At least 40 villagers were killed and several others injured on Saturday as gunmen believed to be Boko Haram members attacked Dalwa-Masuba Village in Damboa Local Government Area of Borno State, security sources and a witness said. The gunmen who stormed the village in large numbers also burnt down virtually all buildings in the village as well as three pickup vans carrying woods to Damboa. A member of the vigilante in Dalwa-Masuba, who spoke to journalists in Maiduguri on phone, said no security personnel had reached the attacked town at the time he was speaking. "We were on patrol somewhere near Damboa when we heard about the attack from some of the villagers who ran from the village", said the source. "We had to drive to the town on our patrol van; we met the entire village on fire, and about 40 persons dead, there were bodies all over the place; three firewood pickups were also set ablaze." The police and the military in Borno are yet to formally confirm the attack, although security sources in Maiduguri, the state capital, said they had been briefed of the attack. Dalwa-Masuba is a farming community 40km away from Damboa Town and about 80km south-west of Maiduguri. The attack follows similar patterns of attacks on communities in Borno by the Boko Haram. The group has continued its attacks and killed thousands of people despite a state of emergency imposed on Borno, Yobe, and Adamawa in May last year. The atrocities of the group, including its kidnap of over 250 teenage, female students in Chibok, Borno State, on April 14, has drawn international attention and condemnation. At a security summit in the French capital, Paris, Saturday, attended by President Goodluck Jonathan, the leaders of Chad, Niger, Benin, and Cameroun, agreed to share intelligence, and co-ordinate action against the group which is based in northeast Nigeria, but has operated somewhat freely in northwest Cameroun, parts of Chad and Niger. A central intelligence platform will be based in Chad, the summit agreed, and will allow all countries involved, including the world powers, to stage a response as necessary. Representatives of the United States, United Kingdom, France, and the European Union, also attended the Saturday's meeting.

Fig. 1: A sample English news report with different place names from Atrocity dataset [2]

For instance, consider the news reports given in Figures 1, 2 and 3 in English, Spanish and Arabic languages respectively. Figure 1 is a report which describes an English atrocity event that occurred in the village of *Dalwa-Masuba*, Nigeria. Moreover, the report also mentions other locations such as Damboa, Maiduguri, Borno, Yobe, Adamawa, Chibok, and Paris. Figure 2 is an Spanish atrocity event about killing a villager in *El Caracol*, Colombia. This report also mentions other locations such as Arauca and Venezuela. Figure 3 is about killing several Palestinian protesters during the opening

978-1-7281-6739-8/19/$31.00 ©2019 IEEE
DOI 10.1109/CIC48465.2019.00017

Tool	Extracted Country	Extracted Locations	Focus Country	Focus Location
Cliff-Clavin	NG, FR, TD, NE, BJ,CM,US,IT	Adamaoua, Benin, Borno, Cameroun, Chad, Chibok, Damboa, France, Niger, Maiduguri, Paris, United Kingdom, United States, Yobe	NG	Borno, Damboa Maiduguri
Geoparser	NG, NE, FR, USA AE,CM,TD,BJ	Adamawa,Benin,Borno,Cameroon,Chad,Chibok,European Union,Faransā, Maiduguri, Niger, Nigeria, Paris, United States, Yobe	-	-
Mordecai	NG	Borno, Cameroun-Gbene, Chibok, Dalwa, Damboa, Komadugu, Yobe, Maiduguri	NG	-
Edinburgh	-	Adamawa, Benin, Borno, Cameroun, Chad, Chibok, Damboa, France, Maiduguri, Niger, Nigeria, Paris, United Kingdom, United States, Yobe	-	-
Stanford CoreNLP	-	Adamawa, Benin, Boko Haram, Borno, Chad, Chibok, Cameroun, Dalwa-Masuba Damboa, France, Maiduguri, Niger, United Kingdom, United States, Yobe	-	-
Mitie	-	Paris, Borno state, Damboa, Maiduguri, Niger, Benin	-	-

TABLE I: Output of focus location extraction from existing tools including Cliff-Clavin, Geoparser, Mordecai, Stanford and Mitie for the given example in Figure 1

Dos hombres armados asesinaron en el corregimiento El Caracol, al comerciante y ganadero de 66 años de edad. Según la denuncia la víctima: Era reconocido por la comunidad araucana por sus servicios y apoyo a los habitantes del corregimiento El Caracol, que se encuentra ubicado a 60 kilómetros al oriente de la capital de departamento de Arauca, en la frontera con la república bolivariana de Venezuela. Gómez Daza, había sufrido dos atentados y recientemente había denunciado ante las autoridades competentes, amenaza a su vida por parte de miembros de la guerrilla del Ejército de Liberación Nacional y paramilitares, en contra de quienes declaró recientemente en audiencia púb lica.

Fig. 2: A sample Spanish news report with different place names from Revista Noche y Niebla[1]

Tool	Extracted Locations
Stanford	Venezuela
Mitie	El Caracol, venezuela, greenarauca

TABLE II: Output of location extraction from Stanford and Mitie for the given example in Figure 2

بدأ حفل تدشين السفارة الاميركية المثير للجدل في مدينة القدس المحتلة الاثنين بعد مواجهات دامية في قطاع غزة مع القوات الاسرائيلية. وبدأ الحفل مع اداء النشيد الوطني الاميركي، قبل ان يبدأ سفير الولايات المتحدة لدى اسرائيل ديفيد فريدمان القاء كلمته. وقال فريدمان ان السفارة تدشن في "القدس، اسرائيل" وسط تصفيق حار من الحضور. واندلعت مواجهات عنيفة الاثنين على حدود اسرائيل مع قطاع غزة ادت الى استشهاد ٤١ فلسطينيا على الاقل، واصابة المئات من الفلسطينيين برصاص الجيش الاسرائيلي قبل ساعات على تدشين السفارة الذي يثير استنكارا دوليا وغضبا فلسطينيا. ويرغب الفلسطينيون في جعل القدس الشرقية عاصمة لدولتهم المنشودة. وكان اعلان ترامب في ٦ كانون الاول(ديسمبر) ٢٠١٧ الاعتراف بالقدس عاصمة لاسرائيل ونقل سفارة بلاده من تل ابيب الى القدس، اثار غبطة الاسرائيليين وغضب الفلسطينيين. واعتبر الكثير من الفلسطينيين قرار ترامب بمثابة استفزاز.-(ا ف ب)

Fig. 3: A sample Arabic news report with different place names from Alghad news agency[2]

Tool	Extracted Locations
Polyglot	Jerusalem, Gaza strip, Israel, United States, Tel Aviv

TABLE III: Translated outputs of location extraction from Polyglot for the given example in Figure 3

ceremony of the US embassy at *Jerusalem* in Arabic. This news report mentions various locations such as the Gaza strip, Israel, United States, and Tel Aviv. For Figures 1, 2 and 3, we say that "Dalwa-Masuba", "El Caracol" and "Jerusalem" are respectively the primary focus locations since the events occurred in these locations. However, other localities associated with the event which are course-grained form the elements of the focus location set.

Even though several geoparsers such as Cliff-Clavin [1], Mordecai [3], and Stanford-CoreNLP [4] have been developed to automatically extract named locations from unstructured English text, location extraction from a text is still a challenging task due to the complexity, diversity, and ambiguity of location information in different languages. However, these tools cannot extract the *focus location* with good accuracy, and most of them cannot differentiate between different locations in the text—i.e. focus locality versus non-focus locality—and are not language agnostic. Following the example in Figures 1, 2 and 3, Tables I, II and III show the output of these different tools for extracting focus location respectively. Clearly, these tools can identify multiple locations mentioned in the news article. Among them, only Cliff-Calvin is able to identify a few focus locations over the English dataset. But it still cannot identify the desired primary focus location. For Spanish and Arabic languages, to the best of our knowledge there is no tool to extract focus location. Furthermore, Stanford NER and Mitie do not support Arabic. Therefore, we utilize *Polyglot* as a NER tool for Arabic news.

One of the main challenges is to identify the primary focus location among the different candidate locations and from news articles in different languages. We address this challenge by using a supervised classification model that leverages contextual patterns in the occurrences of focus locations regardless of the language. Concretely, we first extract candidate locations using a named entity recognition tool and identify the sentences in which they occur. We then extract semantic features from these sentences by using fastText_multilingual model [5] and sentence embedding approaches [6], [7]. Finally, we train a classifier on labeled training instances of different languages and then predict the primary focus location on unlabeled test sentences of different languages. We denote this approach as Primary Focus Location Extraction or *Profile*.

One of the major challenges with the discussed approach is

the lack of suitable labeled data instances for training. In the real world, these labeled instances are not readily available, or may be available scarcely. Traditionally, it is assumed that the training and test data sets used for supervised learning methods are generated from the same data distribution and are monolingual. In practice, this assumption may not be true. In our scenario, true labels of sentences (focus or non-focus) may be only available from monolingual news articles associated with a single news agency or for a small number of news articles. In such cases, these labeled articles may not be a good representative of the population. For example, news articles from different agencies typically have dissimilar linguistic content, vocabularies, writing styles, or type of emotions (e.g., acted, elicited, or naturalistic). Such differences affect classifier performance when employed to predict focus locations in news articles in the wild, and limits the scalability of our approach.

We address this challenge by manually labeling a small number of sentences which creates a sampling bias between the training and test data sets. We then leverage the approach of sampling bias correction by using Kernel Mean Matching (KMM) [8] to estimate the density ratio between the test and training data distributions to appropriately weight each training data instance and then use these instances to train a classifier for prediction of focus location.

The key contributions of this paper are as follows:

- We address the problem of automatically predicting a primary focus location for event-based news reports in different languages by extracting semantic features that aid in capturing patterns of focus location occurrences in various languages, rather than using an external database such as gazetteer[3].
- We propose *Profile* which identifies the primary focus location of political news article in different languages. In particular, we apply supervised learning on a smaller set of biased training data and leverage a well-known bias-correction mechanism to evaluate test data from the same domain to address the label scarcity problem. This approach is language agnostic.
- We empirically evaluate Profile over real-world English, Spanish and Arabic news articles and compare its performance against other available tools.

The rest of the paper is organized as follows. We review related works and present relevant background of geolocation extraction in Section II. We then present our proposed approach in Section III to address the above challenges. Finally, we empirically evaluate our approach in Section IV and conclude in Section V.

II. Background

A. Geolocation Extraction

The field of geolocation extraction collectively involves many different tasks and analyses to be performed over text. The three main tasks among these are: (i) Location named entity extraction [9], [10]; (ii) Location named entity resolution [11]; (iii) Event's location extraction [1], [3].

This paper revolves around the last task, i.e. using geoparsers to extract event location.

Web page geotagging models such as Web-a-where [12] identify all location names using gazetteer, assigns a geographic location and a confidence level to each page, and derive the focus location associated with a web page. Another such framework presented by Silva et al. [13] is used to automatically identify geographic scopes from Portuguese web pages. To locate the geographical entities in web pages, this framework utilizes different external sources such as WHOIS and DNS registrars, and the Portuguese postal codes database. Geoparsers such as Geoparser.io [14] are hosted as web services that identify place names and handle contextual ambiguity among those places. Cliff-Clavin [1] is an open source geoparser which is also hosted as a web service that parses news articles or other documents. It employs context-based geographic disambiguation over organizations and locations extracted using Stanford CoreNLP from the text. It employs a simple frequency-based method to identify the focus places from places mentioned at city, state, and country levels [1]. Mordecai [3] is also an open source geoparser which uses MITIE's NER tool to extract place names from text and then uses gazetteer to identify focus country and all other place names from the text. The Edinburgh Geoparser [15] is yet another geoparser designed to identify occurrences of locations from unstructured text and map them to exact latitude and longitude. NewsStand [16] is another geoparser and geotagger tool. NewsStand extracts the "interesting" phrases that are most likely to be references to geographic locations and other entities by using NER methods. LOCATION phrases are stored as geographic features of the entity feature vector. Then, it uses a Gazetteer to find those geographic features in the entity feature vector that are names of actual locations. It also employs Gazetteer to identify the hierarchical information for each location (i.e. country and administrative subdivisions). After that, it extracts *geographic focus* (or focus location) based on the frequency of the locations in the news. The empirical results of our approach are compared with these competing methods.

Each of the discussed geoparsers capabilities in extracting focus location and focus country are presented in Table IV.

Tools	NER Extraction	Location Extraction	Focus Country	Focus Location
Cliff-Clavin	Stanford CoreNLP	✓	✓	✓ (City, State)
Mordecai	MITIE	✓	✓	✗
Geoparser	NA	✓	✗	✗
Edinburgh	rule-based	✓	✗	✗
NewsStand	LingPipe[4]	✓	✓	✓ (Region)

TABLE IV: Capabilities of different tools in focus country and focus location extraction. Here, ✓ indicates presence and ✗ indicates absence of corresponding capability.

[3]http://www.geonames.org/

It is observed that Mordecai and Cliff-Clavin are the only geoparsers capable in extracting the focus country. Moreover, Cliff-Clavin also extracts focus location on two different levels, i.e. city and state. Almost all of the above mentioned geoparsers are only able to work on English text.
In this paper, we are going to extend our previous research [17], [18] to other languages, such as Arabic and Spanish. Apart from that, due to the lack of label data in languages other than English, we propose an approach to address this challenge in this paper.

B. Multilingual Word Vectorization

Monolingual word vectors are represented in a high-dimensional vector space, such that two contextually similar words are closer in this space [7]. Since these vectors are monolingual, similar words within a language share similar vectors while translated words from different languages do not have similar vectors. To solve this particular problem, Smith et al. [5] presented a framework using Singular Value Decomposition (SVD) to learn a linear transformation (a matrix), which aligns monolingual vectors from two languages in a single vector space without changing any of the monolingual similarity relationships. This model uses Facebook's fasttext word vectors and then uses the Google translate API to translate these words into the 78 languages available. To place all 78 languages in single space, this model aligned every language to the English vectors (the English matrix is the identity).

C. Covariate Shift

An element belonging to a training set is indicated by a subscript tr, while that of a test set is indicated by a subscript te. Each element is indexed by a superscript integer, since a set may contain multiple elements. For instance, $\mathbf{x}^{(i)} \in \mathbf{X}_{tr}$ indicates the i^{th} data instance (array of d covariates) of a training dataset $\mathbf{X}_{tr}$ (a set containing arrays). Also, a hat indicates an estimated value. In general, we use a capital-bold letter to indicate a set of arrays, and a bold letter to indicate an array.

In the case of data classification — a binary focus or non-focus classification in this paper — inequality between the probability distributions of the training and test data sets can be represented in the form of the joint probability distributions $p_{tr}(\mathbf{x}, y) \neq p_{te}(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is the d-dimensional covariate of a data instance with class label y. p_{tr} and p_{te} are the training and test probability distribution respectively. According to Ben-David et al. [19], learning is not possible with bounded error if the two distributions are arbitrarily different. However, this challenge can be addressed by a method that transfer knowledge (model) from training data to test data using instances or feature representation under several assumptions [20].

One such assumption is the equality in class conditional distribution. Concretely, $p_{tr}(y|\mathbf{x}) = p_{te}(y|\mathbf{x})$. Therefore, the inequality in joint probability distribution is attributed to the covariate distribution, i.e., $p_{tr}(\mathbf{x}) \neq p_{te}(\mathbf{x})$. This is known as *covariate shift*. Overall, a correction to the inequality between $p_{tr}(\mathbf{x})$ and $p_{te}(\mathbf{x})$ is provided by computing an importance weight $\beta(\mathbf{x}) = \frac{p_{te}(\mathbf{x})}{p_{tr}(\mathbf{x})}$ for each instance $\mathbf{x}$. This weighted training data set whose data distribution is equivalent to the test data distribution can be used to train a supervised classifier. Various studies have focused on directly estimating the importance weighting function (or density ratio) rather than computing $p_{te}(\mathbf{x})$ and $p_{tr}(\mathbf{x})$ separately. These include Kernel Mean Matching (KMM) [8], unconstrained Least Square Importance Fitting (uLSIF) [21], and Kullback-Leibler Importance Estimation Procedure (KLIEP) [22].

1) Kernel Mean Matching: The main idea in KMM is to decrease the mean distance between weighted training data distribution $\beta(\mathbf{x})p_{tr}(\mathbf{x})$ and the observed test data distribution $p_{te}(\mathbf{x})$ in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{F}$ with feature map $\phi : \mathcal{D} \rightarrow \mathcal{F}$. The mean distance is determined by computing the *Maximum Mean Discrepancy* (MMD), given by

$$\left\| E_{\mathbf{x}\sim p_{tr}(\mathbf{x})}[\beta(\mathbf{x})\phi(\mathbf{x})] - E_{\mathbf{x}\sim p_{te}(\mathbf{x})}[\phi(\mathbf{x})] \right\| \tag{1}$$

where $\|\cdot\|$ is the l_2 norm, and $\mathbf{x} \in \mathbf{X} \subseteq \mathcal{D}$ is a data instance in a dataset $\mathbf{X}$. Here, it is assumed that p_{te} is absolutely continuous with respect to p_{tr}, i.e. $p_{te}(\mathbf{x}) = 0$ whenever $p_{tr}(\mathbf{x}) = 0$. Furthermore, the RKHS kernel h is universal in the domain. It has been proven that under these conditions, minimizing MMD in Equation 1 converges to $p_{te}(\mathbf{x}) = \beta(\mathbf{x})p_{tr}(\mathbf{x})$ [23].

In general, finding desired importance weights by minimizing MMD is equivalent to minimizing the corresponding quadratic program that estimates the population expectation with an empirical expectation. The empirical approximation of MMD (Equation 1) to get the optimal solution for $\hat{\beta}(\mathbf{x})$ is given by

$$\hat{\boldsymbol{\beta}} \approx \arg\min_{\beta} \left\| \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta(\mathbf{x}_{tr}^{(i)})\phi(\mathbf{x}_{tr}^{(i)}) - \frac{1}{n_{te}} \sum_{j=1}^{n_{te}} \phi(\mathbf{x}_{te}^{(j)}) \right\|^2 \tag{2}$$

where $\hat{\beta}(\mathbf{x}) \in \hat{\boldsymbol{\beta}}$. The equivalent quadratic program is as follows.

$$\hat{\boldsymbol{\beta}} \approx \underset{\beta}{\text{minimize}} \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} - \boldsymbol{\kappa}^T\boldsymbol{\beta} \tag{3}$$
$$\text{subject to } \beta(\mathbf{x}^{(i)}) \in [0, B], \forall i \in \{1 \dots n_{tr}\}$$
$$\& \left| \sum_{i=1}^{n_{tr}} \beta(\mathbf{x}^{(i)}) - n_{tr} \right| \leq n_{tr}\epsilon$$

where $\mathbf{K}$ and $\boldsymbol{\kappa}$ are matrices of a RKHS kernel $h(\cdot)$ with $K^{(ij)} = h(\mathbf{x}_{tr}^{(i)}, \mathbf{x}_{tr}^{(j)}) \in \mathbf{K}$, and $\kappa^{(i)} = \frac{n_{tr}}{n_{te}} \sum_{j=1}^{n_{te}} h(\mathbf{x}_{tr}^{(i)}, \mathbf{x}_{te}^{(j)}) \in \boldsymbol{\kappa}$. $B > 0$ is an upper bound on the solution search space, and ϵ is the normalization error. In this paper, we utilize the KMM algorithm for bias correction on training data.

III. Approach

A. Focus Location Extraction

Figure 4 shows an overview of Profile for primary focus location extraction. We first extract candidate focus location

by pre-processing the given news reports, and then utilize a supervised classifier to identify a *primary* focus location among them. In the pre-processing step, since focus locations are mostly mentioned in the first few sentences, we choose a user defined number (denoted by γ) of sentences in each news report. Then, we identify the location named entities in the training news report among these first few sentences using Stanford CoreNLP. Next, we extract the sentence features from select sentences that contain locations. If the sentence includes a focus location, we assign a *Focus* label to it; otherwise, we assign a *Non-Focus* label to it. Finally, we train a binary classifier in a supervised manner using this labeled training data [24], [25].

On the other hand, after the pre-processing step, in the test phase, we assign a label to each sentence in each report using this model. The labels consists of either "Focus" or "Non-Focus". Labeled sentences are called *focus sentences*. Note that more than one focus sentence is included with each news report. Among candidate location names in the collection of focus sentences for each report, we identify the primary focus location using a frequency-based approach to select the focus location. In particular, the frequency-based approach is as follows. We form a histogram of each location detected by NER tools. The location having the highest count is selected as the focus location.

In the next two subsections, we present the features extracted from a text-based dataset. Then we use our learning method to identify the focus locations from an unstructured text.

1) Word Embedding: Our feature extraction algorithm is based on using pre-trained word embedding model from raw text. We utilized the publicly available fastText_multilingual [5] which was built with fastText from Facebook and Google Translate API to align monolingual vectors from two languages in a single vector space. The length of these vectors is 300. We initialize the words that are not present in the set of pre-trained words as zeros. An interesting property of the word embedding is that these vectors effectively encode the semantic meanings of the words in the context. In other words, they are able to represent meaningful syntactic and semantic regularities in a very simple way [7].

2) Sentence Embedding: Our basic sentence feature extraction method follows the Sentence Embedding [6]. We employed this approach because uncommon words are given more weight in the corpus. In other words, common words become less important in the dataset. An alternative approach to find the sentence vector is by computing the mean of the words' vectors in the sentence. We will compare the effectiveness of the Sentence Embedding approach with assigning different weight to each word and the alternative approach empirically in Section IV.

Let c_s be a discourse vector, s be a given sentence, S be a set of sentences and α is a scalar. The discourse vector represents "what is being talked about". Assume that $p(w)$ is the unigram probability of a word in a corpus. Given the discourse vector c_s, the probability of a word w in the sentence s is $p(w|c_s)$.

$$p(w|c_s) = \alpha p(w) + (1-\alpha)\frac{exp(< v_w, \tilde{c_s} >)}{Z_{\tilde{c_s}}} \tag{4}$$

where

$$\tilde{c_s} = \beta c_0 + (1-\beta)c_s, c_0 \perp c_s$$

$c_0 \in \mathbb{R}^d$ is a common discourse vector which serves as a correction term for the most frequent discourse that is often related to syntax, and $Z_{\tilde{c_s}}$ is a normalizing constant given as follows.

$$Z = \sum_{w \in \mathcal{V}} exp(< v_w, \tilde{c_s} >)$$

So, the likelihood for the sentence s is:

$$\begin{aligned} p(s|c_s) &= \prod_{w \in s} p(w|c_s) \\ &= \prod_{w \in s} \left(\alpha p(w) + (1-\alpha)\frac{exp(< v_w, \tilde{c_s} >)}{Z} \right) \end{aligned} \tag{5}$$

where Z is roughly the same as $Z_{\tilde{c_s}}$.

The maximum likelihood estimator (MLE) for $f_w(c_s) = log(p[s|c_s])$ is approximately,

$$\arg\max f_w(\tilde{c_s}) \propto \sum_{w \in s} \frac{a}{p(w)+a} v_w \tag{6}$$

where

$$a = \frac{1-\alpha}{\alpha Z}$$

The MLE is approximately a weighted average of the vectors of the words in the sentence. To estimate c_s, we estimate the direction c_0 by computing the first principal component of $\tilde{c_s}$ for a set of sentences. The final sentence embedding is computed by subtracting the first principle component from $\tilde{c_s}$, since we have to omit the effect of a common discourse vector which is often related to the syntax. More details of this method are described in [6].

The process of feature extraction by using sentence embedding is summarized in Algorithm 1. The inputs of the algorithm are News_Reports, focus_locations, Word_Embedding, and Parameters a and γ. In the first For-loop of the Algorithm (line 1 to 11), we extract set of the locations (loc) by using Stanford CoreNLP as a named entity recognizer (NER) for each news report (line 3), and exclude countries' name from them in line 4. Then, we select the first γ sentences for each news report which contain at least one location name (line 6 to 10). In the next for-loop, we compute the sentence embedding vector ($\mathbf{v_s}$) for each sentence, based on equation 6 (line 12 to 14). For more frequent words w, the weight $\frac{a}{a+p(w)}$ is smaller, so this leads to smaller weights for frequent words. Finally, we compute the first principle component u and decrease it from sentence vector $\mathbf{v_s}$ (line 16 to 18). We trained the SVM classification model using the extracted feature vectors.

We apply the same algorithm during the test process,. However, in the first for-loop, we just use the locations extracted by using Stanford CoreNLP ($loc \leftarrow \mathrm{NER}(\mathrm{News}_i)$).

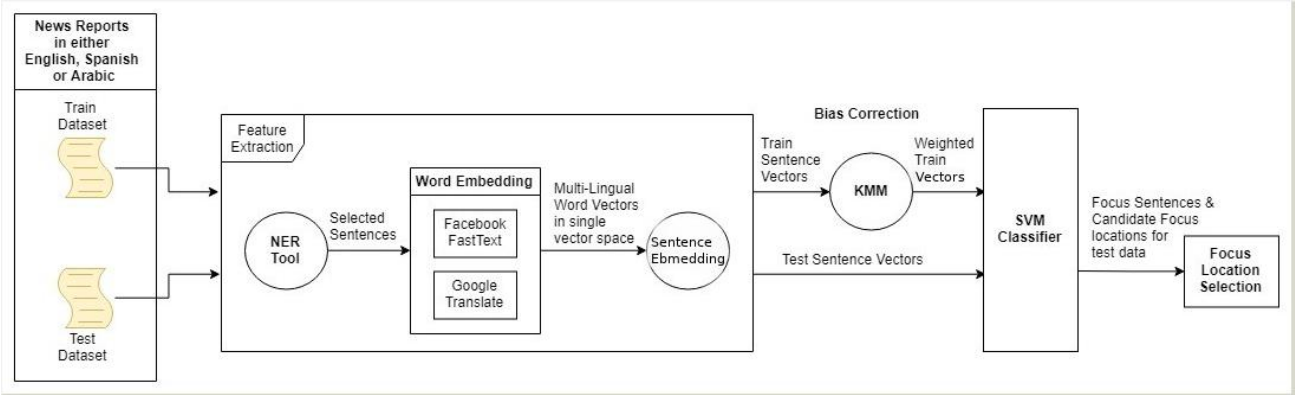


Fig. 4: A high level schema of Profile (Primary Focus Location Extraction).

Then, we classify the feature vectors by using the model. Since there may be more than one Focus sentence per report (i.e., sentence containing potential focus location), we extract the locations from Focus sentences. Next, we use the frequent-based approach to extract the Focus locality. In frequency-based approach, we select the most frequent item in the list. In other words, if we find several sentences from one article with a focus label, the most frequent location name will be a candidate for focus location.

Algorithm 1: Feature Extraction in Profile using Sentence Embedding

Data:
News_Reports $\mathrm{News}_1, ..., \mathrm{News}_N$, focus_locations $\mathrm{Floc}_1, ..., \mathrm{Floc}_N$, Word_Embedding $\{\mathbf{v_w} : w \in \mathcal{V}\}$, Parameter a and γ
Result: Sentence_Embedding $\mathbf{v_s}$

for ***each*** *News_Report* (News_i) **do**
  /* Extract the Location Named Entities for News_i by using a NER tool */
  $loc \leftarrow \mathrm{NER}(News_i) \cup Floc_i$
  $loc \leftarrow \mathrm{loc} \setminus \mathrm{Country\_Names}$
  /* Select the first γ sentences which contain a location in loc. S is a list of these sentences (s). */
  for ***each sentence*** *(s)* **do**
    if $(\#s <= \gamma \; and \; \exists i : \mathrm{loc}_i \in s)$ **then**
      $S \leftarrow S \cup s$
    end
  end
end
for ***each sentence*** s_i *in* S **do**
  $\mathbf{v_{s_i}} \leftarrow \frac{1}{|s_i|} \sum_{w \in S} \frac{a}{a+p(w)} \mathbf{v_w}$
end
/* Compute the first principal component u of v_{s_i} */
for ***each*** *Sentences* s_i *in* S **do**
  $\mathbf{v_{s_i}} \leftarrow \mathbf{v_{s_i}} - \mathbf{u}\mathbf{u}^T\mathbf{v_{s_i}}$;
end

Lang	Dataset	# News Reports	# Sentences
English	Atrocity Event Data	3.6K	40K
	New York Times	1K	103K
Spanish	nocheyniebla	1.5K	8K
	Protest	200	1.5K
Arabic	Alghad News	8.2K	75K

TABLE V: Dataset Statistics

As mentioned earlier, in Section I, we may not have sufficient labeled data to train an unbiased classifier. In such a case, we employ the following approach for bias correction over training data. From the given biased training data, we first perform pre-processing steps by extracting feature vectors. Then over these feature vectors, we apply the bias correction method. Particularly, using KMM we compute instance weight for each training data. This estimates density ratios with the given test data instances. We then train a suitable classifier using the weighted training data in RKHS. This classifier is used to predict focus location over test focus sentences.

IV. Experiments

In this section, we first explain the dataset used to evaluate the proposed method to extract focus locations, and then present the evaluation results while comparing it with the other competing methods.

A. Dataset

The Atrocities Event Data [2] is a collection of recent English news reports on atrocities and mass killings in several locations. Human coders have read the reports and extracted metadata about the events reported. The annotated reports include victims, focus location, and the reports that reported the event. For the training and testing dataset, we excluded the reports that contain multiple events. Moreover, we only select the reports whose locations were correctly extracted by different NER's such as Stanford and MITIE since the performance of NER is beyond the scope of this paper. The original size of Atrocity dataset is about $15K$ reports, and almost $5K$ of them are annotated.

Another English dataset that we used is the New York Times (NYT)[5] news reports dataset. The New York Times Annotated Corpus includes more than 1.8 million articles composed and published by the New York Times between January 1, 1987 and June 19, 2007 with article metadata. Similar to the Atrocity Event dataset, we only select political news articles that contain special keywords such as kill, die, injure, dead, death, wounded and massacre in their title. Although NYT corpus includes location annotations, all of them are not focus locations. Accordingly, we randomly selected 1000 news reports and manually tagged them.

Noche y Niebla[6] is a Spanish dataset with location label (Municipio) for each news report in a PDF format. We developed a PDF extractor to extract news from year 2000 to 2017 in a more structured format[7]. After this process, we removed news reports whose focus locations are not explicitly mentioned in the text. To evaluate the bias correction method, we also asked Spanish-speaking coders to annotate around 200 Spanish protest news articles.

To create an Arabic dataset, we used *Alghad*[8] which is an Arabic news agency. Each news report was initialized with a focus location. We crawled this website to extract news and their focus locations. We used two different categories of news reports, i.e. news of Arabian countries (3K) and World news (5.2K), to test the bias correction method. The overall number of news reports and sentences for English, Spanish and Arabic corpus are given in Table V.

The experiments were conducted on an Intel machine having Core-i7 3.40GHz CPU with 64 GB of RAM, running a standard Ubuntu Linux version 16.04 LTS. We also set $a = 0.1$ and $\gamma = 7$ as inputs for Algorithm 1 as default. We choose $\gamma = 7$ (first seven sentences of any news reports are selected as in input to the algorithm) since we observed that primary focus locations were present in the first 7 sentences of the training set in more that 99% of news reports [17], [18]. We denote this sentence filtering by $\mathrm{Profile}_s$.

Profile uses a support vector machine (SVM) with a Radial Basis Function (RBF) kernel as a base classifier since it supports weighted training data in RKHS. Here, SVM parameter values are $c_{svm} = 1000$ and $\gamma_{svm} = 0.1$.

We compare the performance of Profile with other tools for each of the languages. For English, we used Cliff-Clavin and the frequency-based approach to extract focus locations from Stanford-CoreNLP. Since Stanford-CoreNLP was only developed to identify named entities such as person and location names, it does not distinguish between different levels of location, such as locality and country. Therefore, we modified the Stanford-CoreNLP output and excluded country names from the resulting location names, and then used a frequency-based approach to obtain the most frequent location name as a surrogate for primary focus location. For Spanish, we employed a similar modified Stanford-CoreNLP and Mitie with the frequency-based approach. We utilized Polyglat with the same frequency-based approach for the Arabic dataset. To train the model for each dataset, we randomly picked 60% of articles as training data, and used the remaining 40% of news reports as the test dataset. As shown in Figure 4, Profile first performs the pre-processing steps on both training and test datasets to extract sentences containing potential primary focus location. Then, it extracts related features from the text.

[5]https://catalog.ldc.upenn.edu/ldc2008t19

[6]http://www.nocheyniebla.org/?page_id=399

[7]https://launchpad.net/pdf2xml/+download

[8]https://alghad.com/

Dataset	Method	Accuracy (%)
English (Atrocity)	$\mathrm{Profile}_s$	**71.27**
	Cliff-Clavin	63.75
	Modified Stanford-CoreNLP	60.83
English (NYT)	$\mathrm{Profile}_s$	**64.21**
	Cliff-Clavin	53.65
	Modified Stanford-CoreNLP	36.25
Spanish (Nocheyniebla)	$\mathrm{Profile}_s$	**66.63**
	Modified Mitie	38.44
	Modified Stanford-CoreNLP	29.75
Arabic (Alghad)	$\mathrm{Profile}_s$	**62.41**
	Modified Polyglot	34.43

TABLE VI: Primary focus location accuracy comparison between different methods .

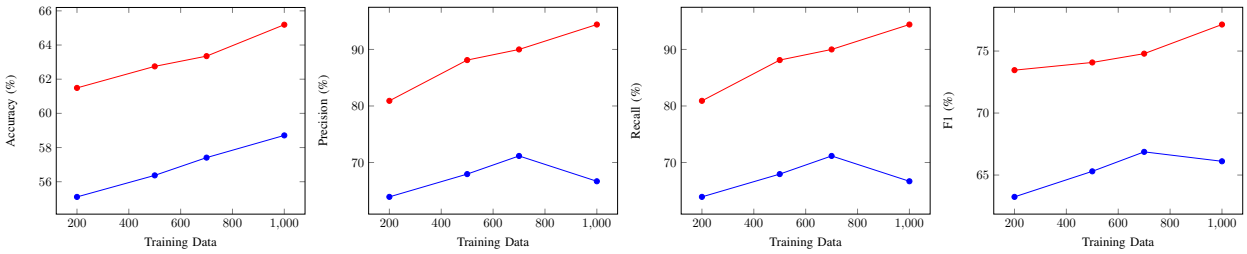
The above experiments assume that the training and test data occur from the same agency or topic. However, a more practical scenario for focus location identification is to study a setting where biased training data is available. We generate a training bias by selecting training data which contains articles from one agency/topic, while the test data contains data from another agency/topic. We utilize the KMM method to obtain an instance weight for each of the training data and build a SVM classifier using the weighted training data. We denote this by $\mathrm{Profile}_s^{\mathrm{KMM}}$. For comparison, we also train another SVM classifier without any weight correction. We denote this by $\mathrm{Profile}_s^{\mathrm{SVM}}$. We then evaluate these classifiers on the same test dataset.

B. Focus Location Extraction Results

We now present the results of Profile where the training and test data occur from the same agency. On average, there are more than five different location names per news report. Note that our goal is to determine the primary focus location while the rest are non-focus locations.

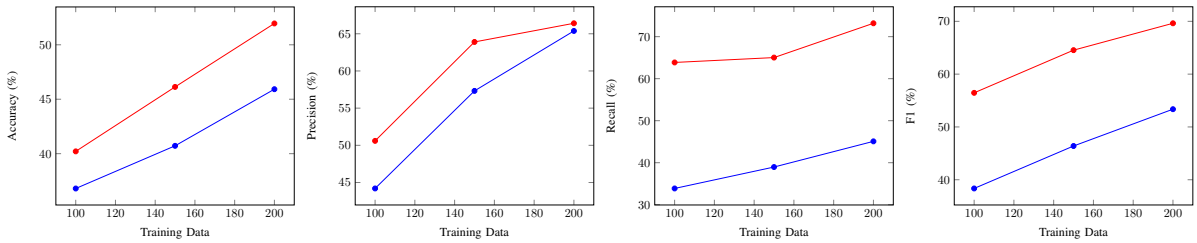
Next, we compare the classification performance of $\mathrm{Profile}_s$ with other existing location estimation approaches. The results are presented in Table VI for English (Atrocity and NYT), Spanish and Arabic datasets. In this table, Cliff-Clavin has better accuracy than modified Stanford CoreNLP for English since it is able to extract the focus country and exclude place names which are not in the focus country. The $\mathrm{Profile}_s$ significantly outperforms modified Stanford CoreNLP and Mitie in Spanish dataset. The proposed approach also surpasses the modified Polyglot approach for an Arabic dataset.

The proposed approach worked better than the existing methods, since we utilize FastText and the sentence embedding model which encoded word semantics and relationships between words in a sentence. Cliff-Clavin can extract locations at a more course-grained level based on the dictionary, and



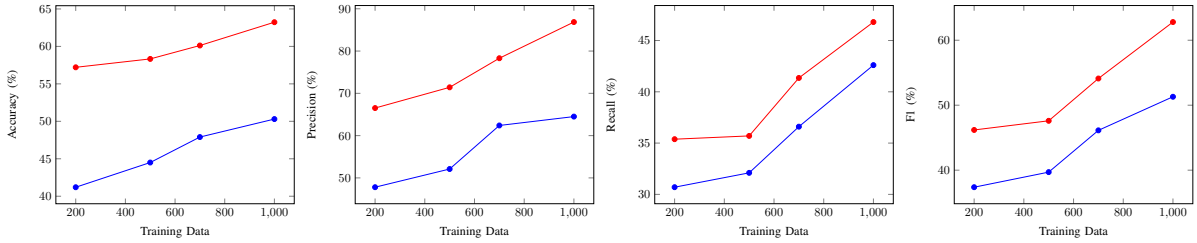
(a) Accuracy (b) Precision (c) Recall (d) F1

Fig. 5: Performance of bias-corrected classifier $\text{Profile}_s^{\text{KMM}}$ with Atrocity dataset as training and NYT dataset as test, compared to a biased classifier $\text{Profile}_s^{\text{SVM}}$.



(a) Accuracy (b) Precision (c) Recall (d) F1

Fig. 6: Performance of bias-corrected classifier $\text{Profile}_s^{\text{KMM}}$ with Spanish Protest dataset as training and Nocheyniebla dataset as test, compared to the biased classifier $\text{Profile}_s^{\text{SVM}}$.



(a) Accuracy (b) Precision (c) Recall (d) F1

Fig. 7: Performance of bias-corrected classifier $\text{Profile}_s^{\text{KMM}}$ with Arabic world news dataset as training and Arabic Arabian news dataset as test, compared to the biased classifier $\text{Profile}_s^{\text{SVM}}$.

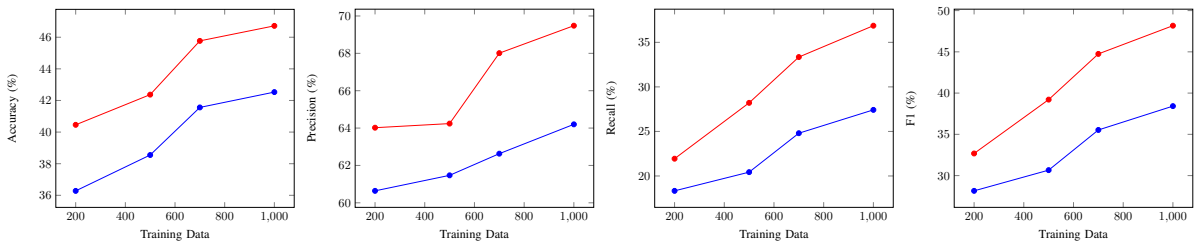
it uses the frequency-based approach to identify the focus locations. As a result, Profile_s outperforms the other methods with 71.27% for Atrocity and 64.21% for NYT. In addition, NER tools are not able to extract focus location at the locality level. To the best of our knowledge, there is not any geoparser that is able to extract focus location from Spanish and Arabic text. Furthermore, most of the popular NER tools are not applicable for Arabic language.

C. Intra-Language Bias Correction Results

Here, we assume the training and test data are from two different publishers/sections in the same language and are related to atrocity news. Figure 5 presents the performance of the $\text{Profile}_s^{\text{KMM}}$ model for focus location extraction in English with Atrocity Event data as the training set and NYT as the test set. Similarly, we also considered Spanish Protest as the training set and *Noche y Niebla* Event Data as the test set. The result is shown in Figure 6 with different sets of randomly selected training data size, following [8]. For Arabic language, we use World news as the training and Arabian news reports as the test dataset. The Arabic result is illustrated in Figure 7.

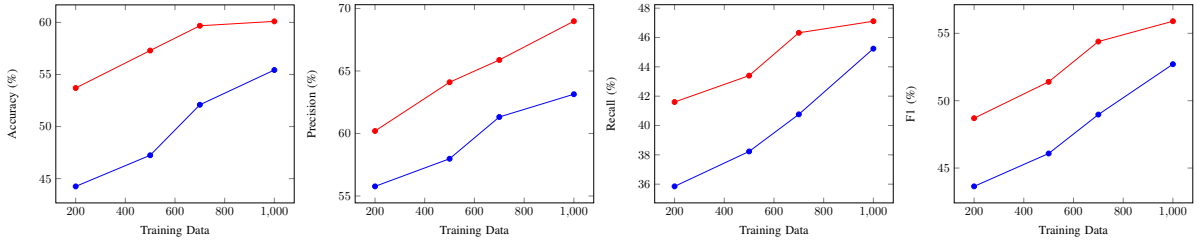
The main conclusions from these three figures are as follows.

- We consistently achieved the best adaptation performance for different training sizes from all experiments on $\text{Profile}_s^{\text{KMM}}$.
- Based on accuracy and precision, we see that $\text{Profile}_s^{\text{KMM}}$ performs similar to the baseline systems. However, $\text{Profile}_s^{\text{KMM}}$ achieves considerably better recall and F1-measure.
- Overall, the $\text{Profile}_s^{\text{KMM}}$ method achieved higher performance than $\text{Profile}_s^{\text{SVM}}$ in all of the experiments.



(a) Accuracy (b) Precision (c) Recall (d) F1

Fig. 8: Performance of bias-corrected classifier $\text{Profile}_s^{\text{KMM}}$ with English Atrocity dataset as training and Spanish Nocheniebla news dataset as test, compared to the biased classifier $\text{Profile}_s^{\text{SVM}}$.



(a) Accuracy (b) Precision (c) Recall (d) F1

Fig. 9: Performance of bias-corrected classifier $\text{Profile}_s^{\text{KMM}}$ with English Atrocity dataset as training and Arabic news dataset as test, compared to the biased classifier $\text{Profile}_s^{\text{SVM}}$.

- Selecting fewer training instances introduced more bias for these domains. As a result, $\text{Profile}_s^{\text{KMM}}$ significantly outperformed the baseline method when the bias is more.
- Finally, the proposed approach works better for English datasets. One of the main reasons is that those datasets are manually selected and labeled. However, this assumption is not true for the Arabic dataset. Since we automatically extracted the news, the data can be noisy or the labels may not have been verified. In addition to this, since we utilized a NER tool at the first step, the performance of the NER tool can affect the performance of Profile.

D. Inter-Language Bias Correction Results

In this section, we train our model with English Atrocity data and then use this model to test on Spanish and Arabic datasets.

Figure 8 presents the performance of the $\text{Profile}_s^{\text{KMM}}$ model which was trained with English Atrocity Event data and tested on Nocheniebla Spanish Event Data to predict the primary focus location. Similarly, we tested the model considering Arabic dataset as the test set. The results are shown in Figure 9 with different sets of randomly selected training data sizes.

The main conclusions from these figures are as follows.

- For accuracy, precision, recall and F1-measure, we consistently achieved the best adaptation performance for different training sizes from all experiments on $\text{Profile}_s^{\text{KMM}}$ when compared to baseline systems.
- Based on these results, we can leverage the availability of labeled data in one language (such as English) to train the model and apply it on other languages to find the focus location, since labeled instances are not always readily available, or may be available scarcely in different languages.

V. Conclusion and Future Work

We showcased and developed a focus location extraction method executable on unstructured text-based news reports from different languages. In this method, we proposed a semantic approach to find the focus location among all the possible locations extracted using the named entity recognition tool. Firstly, we extracted the features using the sentence embedding algorithm. In this algorithm, we encoded the meaning of words and their relationship semantically into a vector regardless of language using the fastText_multilingual model. We then trained a SVM classifier to predict sentences which contain focus or non-focus locations. Finally, we used the proposed method on an Atrocity news event dataset, a subset of New York Times corpus, a Spanish news event dataset from Revista Noche y and an Arabic news event dataset from Alghad.

We applied the domain adaptation technique and conducted experiments over two different domains, since the training and test domain are not always the same. The experiment demonstrated the effectiveness of KMM in extracting focus location when there is bias between different domains.

Our key contribution in this work is extracting the exact focus location at the locality level where an event occurred. The proposed approach works based on the semantic relation-

ship among the words in the sentences and is independent of a geographical dictionary. The performance of our approach exceeds other methods considerably. Furthermore, we proposed the use of a bias correction method to prevent performance loss when the training and test domains are dissimilar.

One of the directions for our future research is extracting primary focus location based on the event. In the current approach, we assume that exactly one event is reported in each news article. Consequently, we extract only one primary focus location from the article. However, this assumption may not be true for all news articles. Therefore, we are going to extract one or multiple event(s) from news articles. The events include different action(s) and some actors. Then, we are going to extract multiple primary focus location(s) related to these events.

Another direction of our future work is developing the bias correction method to improve the results over various domains in news reports. In that case, we can expand our dataset to different domains such as Sport news.

REFERENCES

[1] C. D'Ignazio, R. Bhargava, E. Zuckerman, and L. Beck, "Cliff-Clavin: Determining geographic focus for news," *NewsKDD: Data Science for News Publishing, at KDD 2014*, 2014.

[2] P. Schrodt and J. Ulfelder, "Political instability task force worldwide atrocities dataset," *Lawrence, KS: Univ. Kansas, updated*, vol. 8, 2009. [Online]. Available: http://eventdata.parusanalytics.com/data.dir/atrocities.html/

[3] mordecai, "[online]," *URL: Available: https://github.com/openeventdata/mordecai*.

[4] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[5] S. H. Samuel L. Smith, David H. P. Turban and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *arXiv preprint arXiv:1702.03859*, 2017.

[6] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *International Conference on Learning Representations. To Appear*, 2017.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[8] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Advances in neural information processing systems*, 2006, pp. 601–608.

[9] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[10] A. Ritter, S. Clark, O. Etzioni *et al.*, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1524–1534.

[11] M. Gritta, M. T. Pilehvar, N. Limsopatham, and N. Collier, "What's missing in geographical parsing?" *Language Resources and Evaluation*, pp. 1–21, 2017.

[12] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 273–280.

[13] M. J. Silva, B. Martins, M. Chaves, A. P. Afonso, and N. Cardoso, "Adding geographic scopes to web resources," *Computers, Environment and Urban Systems*, vol. 30, no. 4, pp. 378–399, 2006.

[14] geoparser, "[online]," *URL: Available: https://geoparser.io/*.

[15] B. Alex, K. Byrne, C. Grover, and R. Tobin, "Adapting the Edinburgh geoparser for historical georeferencing," *International Journal of Humanities and Arts Computing*, vol. 9, no. 1, pp. 15–35, 2015.

[16] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, "Newsstand: A new view on news," in *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. ACM, 2008, p. 18.

[17] M. B. Imani, S. Chandra, S. Ma, L. Khan, and B. Thuraisingham, "Focus location extraction from political news reports with bias correction," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1956–1964.

[18] A. K. Gunasekaran, M. B. Imani, L. Khan, C. Grant, P. T. Brandt, and J. S. Holmes, "Sperg: Scalable political event report geoparsing in big data," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 187–192.

[19] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[21] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *The Journal of Machine Learning Research*, vol. 10, pp. 1391–1445, 2009.

[22] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, 2008, pp. 1433–1440.

[23] Y.-l. Yu and C. Szepesvári, "Analysis of kernel mean matching under covariate shift," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 607–614.

[24] Z. Wang, Z. Kong, S. Changra, H. Tao, and L. Khan, "Robust high dimensional stream classification with novel class detection," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 1418–1429.

[25] T. Al-Khateeb, M. M. Masud, L. Khan, C. Aggarwal, J. Han, and B. Thuraisingham, "Stream classification with recurring and novel class detection using class-based ensemble," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 31–40.