Where "where" Matters : Event Location Identification with a BERT Language Model

Hristo Tanev and **Bertrand De Longueville**
Joint Research Centre, European Commission
Ispra, Italy
hristo.tanev@ec.europa.eu
bertrand.de-longueville@ec.europa.eu

Abstract

Detecting event location is a key aspect of event extraction from news and social media. However, this task has not received strong attention recently in comparison to event classification or identifying the event time and the semantic arguments of the event, such as victims, perpetrators, means of action, affected infrastructure, etc. Nevertheless, the location as an event argument plays a crucial role in all event detection applications: conflict detection, health threat monitoring, disaster impact assessment, etc. The method presented in this paper uses a BERT model for classifying location mentions in event reporting news texts into two classes: a place of an event, called *main location*, or another location mention, called here *secondary location* . Our evaluation on articles, reporting protests, shows promising results and demonstrates the feasibility of our approach and the event geolocation task in general.

1 Introduction

Detecting event location from online text sources is a key area of research since the advent of social networks (Intagorn et al., 2010) , (De Longueville et al., 2010). Applications have been developed in fields as diverse as disaster management (De Longueville et al., 2009), (Kongthon et al., 2014), tracking disease outbreaks (Grishman et al., 2002b), or fight against crime (Kounadi et al., 2015). Detecting socio-political events (and in particular, protests) emerged as an important use case (Zhang et al., 2017), which relies on comprehensive, timely and high-quality data that is sometimes not available or it is difficult to be obtained.

Recently, the CASE (Challenges and Application of Automatic Extraction of Socio-political Events from News) series of workshops (Hürriyetoğlu et al., 2021a) have introduced a set of event detection shared tasks and an annotated corpora of protest events, which contains annotations of event places among the other arguments. The CASE initiative significantly boosted the work in the area of socio-political event analysis and gave birth to shared tasks and research works with focus on event location identification, (Giorgi et al., 2021) and (Zavarella et al., 2022).

Formally, geographical place recognition is a sub-category of named entity recognition (NER) (Densham and Reid, 2003). However, it has many particular features: First, geographic names are in the range of millions and unlike names of people and organizations, there are no reliable rules for recognizing these entities by their textual form. Therefore, the first level in recognizing geographic entities is by searching for them in big geographical dictionaries, called *gazetteers*. Second, geographic names can be mismatched with names of people: as an example, let's consider place names like *Washington*, *Georgia*, *Alexandria* and many others. Third, identifying the place names in text is just the first step in recognizing them: disambiguating locations (i.e. which of the many *Paris* is it) and identifying their precise coordinates is even more challenging task (Overell, 2009).

The fourth problem, related to location analysis is recognizing the semantic role of the location mentions. Currently, very little work is dedicated to this important problem: Our paper aims at filling this gap, by applying the latest advances in Natural Language Processing (Devlin et al., 2018), leveraging large language models and the knowledge encoded in them, to recognize locations, where events happen, distinguishing them from other location mentions.

Our approach is designed to be used as an integral part of an automated process for Event Extraction: In particular, we aim at linking protest events from news articles to the locations where they took place. A classical problem of such location identification is the fact that apart from the locations of the main events reported in the news, here called *main locations*, many more places are usually referred

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 11–17, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_002

to in the text, called here *secondary locations*. Typically, a news article focuses on one main event, which however is related to various reported real or possible happenings, which took place before or after the main one. Each event also has an elaborated structure and may feature different semantic arguments, among them places, as well as sub-events and larger events, which encompass it. Conversely, locations may be used to define the places of the events, as well as to address the origins and affiliations of people and organizations (“refugees from Syria", “the mayor of Brussels"),

To answer the event location detection challenge, we proposed an approach which uses a BERT (Devlin et al., 2018) model for text classification; it classifies each location as the place of an event on which the news article is focused (main location) or as a secondary location mention (places of secondary events or location mentions, which are not event places). Our model uses only lexical context and the position of the sentence in the article. However, our approach makes use of the implicit semantic knowledge about the similarities of the words and their relations, encoded in the BERT model. In this way, we avoid using semantic features and other text pre-processing, relying entirely on the semantic knowledge, encoded in BERT. Moreover, recent research (Muller et al., 2022) aims at transferring BERT models across languages , potentially bringing making our our approach multilingual.

2 Related work

Earlier work on event extraction, such as (Humphreys et al., 1997) and the REES system (Aone and Ramos-Santacruz, 2000) use syntactic patterns for detecting locations and other event arguments. Similarly, one of the first disease outbreak systems, PULSE (Grishman et al., 2002a), makes use of syntactic clues and proximity to essential event arguments, such as disease names, to select the outbreak locations. Some recent approaches for event location detection like (Giorgi et al., 2021) also makes use of proximity of the location to specific event arguments.

These linguistic approaches, although having a reasonable precision, are limited in their application, since syntactic patterns and clues require a significant amount of expert knowledge and efforts and strongly depend on the event classes, which are being considered in the event extraction system. Moreover, linguistic approaches cannot efficiently exploit big data repositories, i.e. corpora and public event data bases, such as ACLED (Raleigh et al., 2010), Global Terrorism Database (LaFree and Dugan, 2007), and others, which has recently emerged.

In contrast, Machine Learning (ML) models can significantly benefit from such data: A recent ML work on event geolocation, based on an existing event data set (ICEWS (Ward et al., 2013)) is presented in (Lee et al., 2019). Their work is similar to the approach presented in this paper. However, they use semantic pre-processing of the text by annotating each event-specific keyword and expression: event trigger verbs and nouns (e,g, *breaking into*), actors (e,g, *Ukraininan soldiers*), temporal expressions and others. This work reports 75% accuracy for detecting the main event location in a protest event data set. They use Support Vector Machines (SVM), Neural Network and Random Forests, all methods delivering similar results.

Another work which relies on training a classifier for event location detection is presented in (Imani et al., 2017). They use SVM classifier and word embeddings for identifying the sentences likely to contain the main event location. Then they extract from them the most frequent location.

Similartly, (Halterman, 2019) proposes a Convolutional Neural Network, which finds the main event location. They have manually created a data set of 8’000 sentences, containing information about military offensives in the Syrian war. The event geolocation accuracy they achieve is around 84%.

3 Protest events and their locations

Protest are socio-political events, which include rallies, protests, marches, strikes, riots, violent disorders and civil unrest. Each socio-political event assumes action by a large group of people. In particular, the protest events express disapproval and oppose to concrete actions or policies of governments, administrators, parties, institutions or companies. The definition of protest event, given in (Makarov et al., 2015) is:

A protest event is open to the public, politically motivated and not institutionalised as opposed to e.g. elections.

In some cases protest actions pose concrete demands, e.g. lowering taxes or raising wages. On certain occasions, these events attempt to focus the public attention on causes, such as minority rights, peace in war zones, environmental problems, etc. These events may include spontaneous violent actions, mass violence against people, vehicles and infrastructure. Such actions may manifest characteristics of crime or event small-scale armed conflicts, when armed opposition to the police takes place.

In order to better understand the dynamics of the protest events and their relation with various geographic locations, we have manually analyzed a small set of news, identifying the main and the secondary events and their related locations.

Our analysis found four basic types of location mentions:

- The place of the reported protest, *the main location*, “Farmers staged a protest in *Santa Fe* province on Tuesday ", “they hacked Sabata Petros Chale to death in *Marikana West* , allegedly over the allocation of low cost houses". Main locations can be reported using several levels of accuracy, for example mentioning the country, the district, the city and the place inside the city, e.g. “Clashes erupted in *Dalian*, *Liaoning*" , resulting in several location mentions, referring to a single event. Also, in some cases, more than one main event can be reported, causing mentioning of more than one main location.

- The place of the event which is the cause for the protest - “The incident came about as protests and riots formed in cities across the country following the killing of George Floyd in *Minneapolis* "; “A demonstration against supplying *Ukraine* with weapons for war with *Russia* attracted 10,000 people on Saturday" Such locations we consider *secondary*.

- Another source of secondary location mentions are the populated places from where the protesters come, also their national origin - “Farmers from the nearby states of *Punjab*, *Haryana* and *Uttar Pradesh* began arriving by tractors and on foot at the outskirts of New Delhi last week, where they blocked roads and set up makeshift camps"

- Locations related to response actions and consequences: places of police block, places of blocked traffic, countries reacting towards the event, and places where politicians or organizers make statement about the main event (“press conference with the French Prime Minister in Paris about the protests across the country"). Although these locations may be important for the dynamics of the event reported, they are still considered *secondary locations*.

Let’s consider as an example, a news article fragment describing a protest in Oslo:

“Dozens of activists, including Greta Thunberg of neighboring *Sweden*, blocked the entrance to the energy ministry in *Oslo* Monday to protest a wind farm they say hinders the rights of the Sami Indigenous people to raise reindeer in *Arctic Norway*"

In this fragment three locations are mentioned, while only one, i.e. *Oslo*, is the main location. The other locations mentions , (*Sweden* and *Arctic Norway*), are secondary ones. The first relates to the origin of one of the prominent protesters (Greta Thunberg) and the second is the place, where the cause for the protest is located: a wind farm in *Arctic Norway*.

Our analysis shows that the complexity of the events, described in the news, not only the sociopolitical ones, has its impact on the location references: one happening can trigger mentions of multiple related events and people, and the corresponding locations, related to them. In the case of protest actions, the event which is a cause for them is frequently mentioned along with its place. Moreover, the effects of the protest on the people and the urban environment: blocked traffic, police actions and similar, bring in the text additional locations.

In some sense, this is in agreement with Davidson’s view on the event semantics (Davidson, 1969), for whom the cause and effect constitute important characteristics of the event phenomena.

4 Approach

The approach we propose for geolocating events belongs to the class of Machine Learning approaches, it is similar in spirit to the work of (Halterman, 2019). We, however, chose to use a BERT classification model, since it provides the necessary level of abstraction by encoding the texts into a semantic space, trained on millions of documents. In this way, we avoided the feature abstraction phase, which is part of all the other ML approaches for geocoding, cited so far.

In order to train and evaluate our approach, we used a corpus of protest news, reporting various types of protests in India and China. (Hürriyetoğlu et al., 2021b). In this corpus the annotated event locations are main locations. Moreover, we have additionally annotated the secondary location mentions, which were not annotated in the corpus, using the Mordecai open source software (Halterman, 2017). In this way, we obtained a corpus with main and secondary locations.

A sample annotated sentence from the corpus is shown below:

"$India_{[main]}$: $NewDelhi_{[main]}$, Thu May 30 2013 , 22:07 hrs Activists of Youth Indian National Trade Union Congress (INTUC) protest against recent Naxal attack on Congress leaders , in $Raipur_{[secondary]}$ on Thursday . "

4.1 Generating location windows

We used the following procedure to extract location-specific data from the annotated corpus:

1. We found each main or secondary location mention.

2. We masked each location mention with a placeholder token *EVENT_PLACE* (both for main or secondary locations) and extracted a *location window* of maximum of twenty one tokens from the same sentence: maximum of ten tokens before and after the placeholder without crossing the sentence boundaries.

3. After several experiments, we have found out that the BERT model is sensitive to the exact position of the location place holder, therefore, for the shorter windows we have artificially inserted before and after a series of filler tokes (BEGIN before the window or END after it), so that the length of the window is always twenty one tokens and EVENT_PLACE is in the center of the window.

4. In order to account for the position of the sentence inside the article, we inserted the position of the sentence in front of every location window. Some smaller scaled experiments, not reported here, showed to us that the number of the sentence slightly contributes to the accuracy of the model.

5. Finally, we assigned a label to each of the location windows which shows if it was a *main location* (the place of the event annotated manually) or a *secondary location* (any other location mention, annotated by the Mordecai tool).

Table 1 shows several samples of location windows with the sentence position and the EVENT_PLACE placeholder. For clarity, we do not show the BEGIN and END filler tokens. Each window is labeled as a main location or a secondary one.

4.2 Fine-tuning the BERT model

Location windows were used to fine-tune a FastBERT model (Liu et al., 2020), thus obtaining a large language model which classifies a geolocation as a main or a secondary location mention, using only its location window.

The FastBERT was chosen because of its speed of performance, which allowed us to experiment with multiple data splits in reasonable time. Moreover, the speed of the model is crucial, when applying it in real-world settings: The FastBERT speed can be flexibly adjusted in the classification phase. Moreover, this model adopts a unique "self distillation mechanism" at fine-tuning, further enabling a greater computational efficacy with minimal loss in performance.

5 Experimental set up

In our experiments we used the corpus of protest events with already annotated locations, enriched with automatic location identification from Mordecai, as explained in the previous section.

Following the procedure for extraction of location windows (Table 1) from the annotated corpus, we have obtained an experimental data set of 829 main location windows, considered here as positive instances, and 472 secondary location ones, considered as negative ones.

From this data we have performed a cross-validation, generating 10 random train/test data splits, each containing 66% location windows for training and 34% for test.

We fine-tuned the FastBert model on the training set of each data split and evaluated the performance of the model on the test set.

In order to evaluate the difficulty of the location classification task, we introduced also a simple baseline *First sentence*, which considers a location

Table 1: Data sample. Main and secondary location text windows.

Location window	Main or second.
1 The house of a PDP MP was torched in south EVENT_PLACE .	Main
0 AM The clash between police and the local people in EVENT_PLACE .	Main
6 In EVENT_PLACE district, about 25-30 Maoists attacked the premises of	Main
5 midnight , they set fire to the tower in EVENT_PLACE police station area .	Main
2 The agitation was organized by the EVENT_PLACE district unit of the BJP .	Main
5 Passengers to the EVENT_PLACE airport did not have much of a problem .	Secondary
3 The march was intercepted at the EVENT_PLACE .	Secondary
7 thanks to the providential arrest of a terrorist in EVENT_PLACE	Secondary
0 Seers protest arrest at EVENT_PLACE police station 17th January	Secondary

mention to be a main location (positive), only if it appears in the first sentence of the news article. We also compared the performance of our BERT model to the performance of an SVM, classifier, which uses the Radial Base Kernel Function (RBF) with C parameter set to 1.

We performed 2 runs of the SVM model: In the first run we used bag-of-word vectors, where each dimension corresponds to a word and its value is the number of the word appearance in the text window. In the second run of the SVM model we used Word2Vec Google News vectors (Church, 2017), which are Word Embeddings with 300 dimensions, pretrained on 3 billion Google News Texts.

6 Evaluation

We have calculated precision, recall, the F1 measure and the accuracy of the FastBERT, the two SVM models, bag-of-words (BoW) and Google News Word2Vec (W2V), and the First sentence baseline on the test set of each of the 10 data splits.

In Table 2 we report the average FastBERT performance across the 10 splits, as well as the average performance of the SVM models and the baseline First sentence.

Clearly, FastBERT significantly outperforms the baseline First sentence, especially as a recall, F1 measure and accuracy. Notably, the recall of FastBERT is more than twice the recall of the baseline: This shows the importance of the model for identifying main event locations, which can frequently be mentioned after the first sentence.

Compared to the SVM BoW and SVM W2V, our method showed significantly better accuracy with respect to the two SVM models: 0.73 vs 0.64 for SVM BoW and 0.65 for SVM W2V. The $F1$ measure of BERT and the SVM models are comparable, still BERT outperforms the two SVM models with 0.02 and 0.03.

The standard deviation of the F1 of FastBERT across all the 10 splits is $s = 0.03$. This shows that our evaluation was reliable and the results do not depend strongly on the data split.

Our evaluation shows that BERT outperforms two state-of-the-art machine learning models and a baseline for detecting event locations.

Although not directly comparable, the results we achieved are similar in terms of accuracy to the results reported by (Lee et al., 2019) on a different data set of protest events: The best accuracy they achieve is 0.75, using SVM. Their approach, however, uses a significant amount of semantic and morphological pre-processing. In contrast, we entirely relied on the semantic knowledge encoded in the BERT model. This is a clue that the BERT models could decrease the need of extensive feature engineering and provide a basis for non complex identification of event arguments.

7 Conclusions

The objective of this paper was to validate an approach, based on the use of a large language model (FastBERT) to leverage context and semantics in the task of detecting primary (main) event locations. In this process we completely avoided complex feature engineering and linguistic pre-processing. We achieved encouraging results, outperforming an heuristic baseline and SVM classifiers based on bag of words and word embedding vectors.

In this work we focused on protest events, since they are important measure for the level of political discontent in the society and provide a basis for conflict prediction. Other socio-political events, such as armed conflicts, manifest similar problems when analysing their spacial dynamics. In this line of thought, locations are important parameters

Table 2: Evaluation and comparison of BERT with SVM and a baseline

Model	Precision	Recall	F1 measure	Accuracy
FastBERT	0.75	0.86	0.80	0.73
SVM BoW	0.64	0.99	0.78	0.64
SVM W2V	0.65	0.95	0.77	0.65
Baseline First sentence	0.68	0.34	0.46	0.40

for each news report. Moreover, distinguishing main location mentions from secondary ones is an important and challenging task. Therefore, our work has larger scope and applicability which goes beyond the protest events.

The question of performance of such approach for less resourced languages should be tackled. Being multilingual by design is of paramount importance for many Automatic Event Detection applications. The promise of last generation models to transpose learning efficiently from one language to another is in this view a strong incentive to further invest in their use. In this perspective training, testing and evaluating the latest large language models with multi-lingual annotated event location corpora is a relevant research direction in the context of automated location analysis in news and social media streams.

References

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Sixth applied natural language processing conference*, pages 76–83.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Donald Davidson. 1969. *The Individuation of Events*, pages 216–234. Springer Netherlands, Dordrecht.

Bertrand De Longueville, Gianluca Luraschi, Paul Smits, Stephen Peedell, and Tom De Groeve. 2010. Citizens as sensors for natural hazards: A vgi integration workflow. *Geomatica*, 64(1):41–59.

Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. 2009. " omg, from here, i can see the flames!" a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*, pages 73–80.

Ian Densham and James Reid. 2003. System demo: A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 79–80.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, et al. 2021. Discovering black lives matter events in the united states: Shared task 3, case 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002a. Information extraction for enhanced access to disease outbreak reports. *Journal of biomedical informatics*, 35(4):236–246.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002b. Real-time event extraction for infectious disease outbreaks. In *Proceedings of Human Language Technology Conference (HLT)*, pages 366–369.

Andrew Halterman. 2017. Mordecai: Full text geoparsing and event geocoding. *The Journal of Open Source Software*, 2(9):91.

Andrew Halterman. 2019. Geolocating political events in text. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 29–39.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021a. Challenges and applications of automated extraction of socio-political events from text (case 2021): Workshop and shared task report. *arXiv preprint arXiv:2108.07865*.

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021b. Cross-context news corpus for protest event-related knowledge base construction. *Data Intelligence*, 3(2):308–335.

Maryam Bahojb Imani, Swarup Chandra, Samuel Ma, Latifur Khan, and Bhavani Thuraisingham. 2017. Focus location extraction from political news reports with bias correction. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1956–1964. IEEE.

Suradej Intagorn, Anon Plangprasopchok, and Kristina Lerman. 2010. Harvesting geospatial knowledge from social metadata. In *ISCRAM*.

Alisa Kongthon, Choochart Haruechaiyasak, Jaruwat Pailai, and Sarawoot Kongyoung. 2014. The role of social media during a natural disaster: A case study of the 2011 thai flood. *International Journal of Innovation and Technology Management*, 11(03):1440012.

Ourania Kounadi, Thomas J Lampoltshammer, Elizabeth Groff, Izabela Sitko, and Michael Leitner. 2015. Exploring twitter to analyze the public's reaction patterns to recently reported homicides in london. *PloS one*, 10(3):e0121848.

Gary LaFree and Laura Dugan. 2007. Introducing the global terrorism database. *Terrorism and political violence*, 19(2):181–204.

Sophie J. Lee, Howard Liu, and Michael D. Ward. 2019. Lost in space: Geolocation in event data. *Political Science Research and Methods*, 7(4):871–888.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. FastBERT: a self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6035–6044, Online. Association for Computational Linguistics.

Peter Makarov, Jasmine Lorenzini, Klaus Rothenhäusler, and Bruno Wüest. 2015. Towards automated protest event analysis. *New Frontiers of Automated Content Analysis in the Social Sciences*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2022. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *TALN 2022-29° conférence sur le Traitement Automatique des Langues Naturelles*, pages 450–451. ATALA.

Simon E Overell. 2009. *Geographic information retrieval: Classification, disambiguation and modelling*. Ph.D. thesis, Imperial College London (University of London).

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing gdelt and icews event data. *Analysis*, 21(1):267–297.

Vanni Zavarella, Hristo Tanev, Ali Hürriyetoğlu, Peratham Wiriyathammabhum, and Bertrand De Longueville. 2022. Tracking covid-19 protest events in the united states. shared task 2: Event database replication, case 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 209–216.

Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, and Jiawei Han. 2017. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 595–604.