RiCL Research in Corpus Linguistics

The FGLOCTweet Corpus: An English tweet-based corpus for fine-grained location-detection tasks

Nicolás José Fernández-Martínez
Catholic University of Murcia / Spain

Abstract – Location detection in social-media microtexts is an important natural language processing task for emergency-based contexts where locative references are identified in text data. Spatial information obtained from texts is essential to understand where an incident happened, where people are in need of help and/or which areas have been affected. This information contributes to raising emergency situation awareness, which is then passed on to emergency responders and competent authorities to act as quickly as possible. Annotated text data are necessary for building and evaluating location-detection systems. The problem is that available corpora of tweets for location-detection tasks are either lacking or, at best, annotated with coarse-grained location types (e.g. cities, towns, countries, some buildings, etc.). To bridge this gap, we present our semi-automatically annotated corpus, the *Fine-Grained LOCation Tweet Corpus* (FGLOCTweet Corpus), an English tweet-based corpus for fine-grained location-detection tasks, including fine-grained locative references (i.e. geopolitical entities, natural landforms, points of interest and traffic ways) together with their surrounding locative markers (i.e. direction, distance, movement or time). It includes annotated tweet data for training and evaluation purposes, which can be used to advance research in location detection, as well as in the study of the linguistic representation of place or of the microtext genre of social media.

Keywords – location detection; locative references; fine-grained locations; tweets; corpus for training and evaluating models

1. INTRODUCTION

Location detection is an important task in Natural Language Processing (NLP) whereby locative references mentioned in texts are identified and extracted for a variety of practical purposes (Middleton *et al.* 2018; Purves *et al.* 2018). This task has recently been applied to microtext genres such as tweets which, due to their brief and informal nature, contain many non-standard forms that challenge the performance of current NLP systems which are typically trained on more formal genres such as news (Baldwin *et al.* 2013; Eisenstein 2013). Hence, there is an increasing need to focus on building and using corpora based on social media microtexts.

Location detection from social media microtexts has wide-ranging practical applications: from natural or human-made disaster detection and tracking in floods, earthquakes, storms, civil unrest, war, crime, etc. (Vieweg *et al.* 2010; Crooks *et al.* 2013; Imran *et al.* 2014; Jongman *et al.* 2015; Martínez-Rojas *et al.* 2018; Siriaraya *et al.* 2019; Zhang *et al.* 2019), health surveillance and disease tracking (Eke 2011; Dredze *et al.* 2013), for example, the COVID-19 pandemic (Singh *et al.* 2020), to marketing and advertising purposes (Mourad *et al.* 2019), or traffic incident detection, road traffic control and/or traffic congestion (Ahmed *et al.* 2019; Das and Purves 2019; Gonzalez-Paule *et al.* 2019; Khodabandeh-Shahraki *et al.* 2019). In this regard, the extraction of fine-grained geospatial information from social media microtexts plays a key role in intelligent systems for crisis management services to raise emergency situation awareness from crisis-related events where the location dimension proves essential to understand their impact: where an incident happened, where people are in need of help and/or which areas have been affected (Vieweg *et al.* 2010; Crooks *et al.* 2013; Imran *et al.* 2014). Such information could potentially help save lives and/or prevent further damage to environmental or urban areas in emergency- and crisis-related contexts.

Corpus building in this area helps train location-detection systems in supervised probabilistic-based models, typically based on machine learning or deep learning, or develop rule-based systems and assess their performance, with a view to replicating their performance in real-life contexts. The problem is that (i) most tweet corpora are not available for public use, impeding any replication or future development, and (ii) that corpus development in this area has extensively focused on annotating coarse-grained location types such as geopolitical entities (e.g. countries, cities or towns), leaving aside many other toponymic entities that are equally, if not more, important in crisis-related scenarios, such as points of interests, natural landforms and traffic ways. Also, information related to distance, direction or time that may accompany such entities is not tagged, losing again the granularity needed for emergency-based services.

To address these issues, we present the *Fine-Grained LOCation Tweet Corpus* (FGLOCTweet Corpus), which has been semi-automatically built using our linguistically aware location-detection system LOcative Reference Extractor (LORE) for its processing and annotation (Fernández-Martínez and Periñán-Pascual 2021a), including the anonymization of users' references, and supervised error revision. The corpus integrates English tweets with annotated coarse- and fine-grained locative references from real-life

situations for the development and evaluation of location-detection systems with an interest in a greater diversity, variety and semantic granularity of the location types. We may release the corpus upon request[1] for its use in location-detection research development or for linguistic inquiry of the microtext genre and the representation of spatial knowledge in English.

The present article is structured as follows. First, we briefly examine related work in tweet location detection paying special attention to the corpora used, their characteristics and their availability. Then, we provide the methodological steps in building and annotating our corpus. Finally, we discuss the practical uses and applications for research practitioners and conclude with some future research directions.

2. Brief overview of the literature

We provide here, in chronological order, some of the major contributions in corpus building for tweet-based location detection tasks. Most authors have built their own corpus containing thousands of tweets focusing on geopolitical entities (e.g. cities, towns and countries), and have typically restricted themselves to specific areas or crisis-related events (Inkpen *et al.* 2017; de Bruijn *et al.* 2018). However, most of these self-compiled corpora are unavailable for public use, and they contain, most of the times, coarse-grained locative references only. Other problems relate to the use of different corpus annotation standards, which aggravates the reutilization of such resources.

Inkpen *et al.* (2017) built, for their probabilistic-based location-detection system, a corpus of 6,000 semi-automatically annotated tweets containing 4,369 mentions of coarse-grained locations (i.e. cities, provinces and states) from the US and Canada.[2] The building process consisted of two phases: first, a simple matching step was performed using the *GeoNames* database (Ahlers 2013) to match names of locations from the US and Canada together with their location type and, then, a manual filtering process by expert annotators was conducted to revise errors or include other missed entities. Their corpus missed richer locative reference types such as points of interests, streets or

[1] According to *Twitter*'s Privacy and Developer policies, "[…] all developers may provide up to 50,000 public Tweets Objects and/or User Objects to each person who uses your service on a daily basis if this is done via non-automated means (e.g., download of spreadsheets or PDFs)." (Developer Policy – *Twitter* Developers 2021). This means that we can only share these tweets upon users' requests for non-for-profit purposes.

[2] Available at https://github.com/rex911/locdet (5 July, 2021).

highways. Likewise, other important locative markers were ignored (e.g. distance markers such as *n kilometres away from X*), and it only focused on US and Canada entities, leaving aside many other geographic areas of the world.

De Bruijn *et al.* (2018) compiled 2,785 flood-related tweets, manually tagging geopolitical entities such as countries, cities, towns and villages from different parts of the world up to a number of 2,079 locative references mentioned in those tweets, by using a matching algorithm and the *GeoNames* database.[3] Since only geopolitical entities were labeled, the corpus lacks a great deal of fine-grained locative references and potential locative markers.

The most famous available corpus for location-detection purposes is the *GeoCorpora*, built by Wallgrün *et al.* (2018).[4] *GeoCorpora* contains 6,711 tweets of a variety of crisis-related events using keywords as diverse as *floods*, *riots*, *tornados*, *flu*, *violence*, etc. with their correspondingly mentioned locative references, a unique ID and geographic coordinates obtained from the *GeoNames* database, whenever available. Geographers were used to tag and revise the annotation of place names. Since only tweet IDs are provided to retrieve the tweets, it is possible that many may have been deleted by now. The locative types considered in the annotation of the corpus were mostly towns, cities, states and countries, as well as some natural landforms and a few traffic ways (e.g. street names and addresses). Sometimes, location-indicative nouns (e.g. lake, hill, county or state) were tagged as part of locative references. However, the corpus lacks a great deal of location types, and locative markers are not considered.

Hu and Wang (2020) obtained, preprocessed and annotated 1,000 tweets out of a very large corpus of 7,041,866 tweets collected in the event of the Hurricane Harvey in the US in 2017.[5] They performed a study of the location types mentioned in those tweets, differentiating the following: addresses, street names, highways, exit of highways, roads, natural landforms, buildings and geopolitical entities of different types. They also assessed general-domain entity recognizers and found that they fail at detecting traffic ways tremendously. To this date, this is the only released corpus providing a number of easily accessible annotated fine-grained locative references. However, its focus is on a

[3] The code is available at https://github.com/jensdebruijn/TAGGS (5 July, 2021.), but the corpus is not publicly available.
[4] Available at https://github.com/geovista/GeoCorpora (10 September, 2021.).
[5] Available at https://github.com/geoai-lab/HowDoPeopleDescribeLocations (10 September, 2021.).

particular event in a specific area, thus limiting its application to any other crisis-related event in other parts of the world.

Recent research highlights that the task of location detection in social microtexts is not a solved task (Wang and Hu 2019). Specifically, it has been mentioned that there is an ever-growing need for detecting fine-grained locations (street names or the names of parks and monuments), as well as developing corpora of social media microtexts for training and evaluating models with fine-grained locative references (Gritta *et al.* 2018). Considering such limitations in the state of the art, the FGLOCTweet Corpus is intended to address this gap by providing a great number of tweets with annotated fine-grained locative references from a variety of incidents and crisis-related events from all over the world.

3. Methodology

Corpus building involved a series of steps that will be discussed in further detail later in this section. These include corpus compilation, corpus preprocessing and corpus annotation (Rayson 2014: 33). All these stages were performed with semiautomatic techniques (regular expressions, automatic tagging, and manual revision) that greatly facilitated the construction of the corpus. Also, the corpus was built with methodological corpus-based principles in mind in each of the different steps, including size, representativeness and balance, as well as practicality (Reppen 2010). To be more specific, the FGLOCTweet Corpus was built with the aim of capturing as wide a variety of locative references as possible from as many different real-life incidents from as many places in the world. A total of 9,405 tweets with their corresponding tagged tokens seemed to be the sweet spot for locative-detection tasks, since the size of such a specialized corpus does not need to be particularly large but surely needs to be sufficiently representative, in accordance with the corpus size found in the literature of location detection.

In the case of our corpus, which was built for NLP applications, it was also important to consider consistency in the annotation of data or, in other words, that a set of guidelines was followed for the correct training of our model and also for a correct evaluation of the model (Zinsmeister *et al.* 2009: 764). This meant using the same set of part-of-speech (POS) tags. As for the annotated locative references, this implied that these

adhere to i) morphological, structural and semantic criteria involving a definition of what a locative reference is, and ii) the degree of geospatial granularity needed for real-life emergency-based applications. We define a locative reference as a proper noun that designates a geographically locatable spatial point, morphologically realized with full names, abbreviations, acronyms or a given combination of all of them. Structurally speaking, they are either simple or complex, depending on the number of lexical and/or phrasal elements accompanying the proper noun(s). This is illustrated in Figure 1, where an asterisk is used to mark optionality and double asterisk refers to the optional presence of locative markers either at the beginning or at the end of the locative reference.

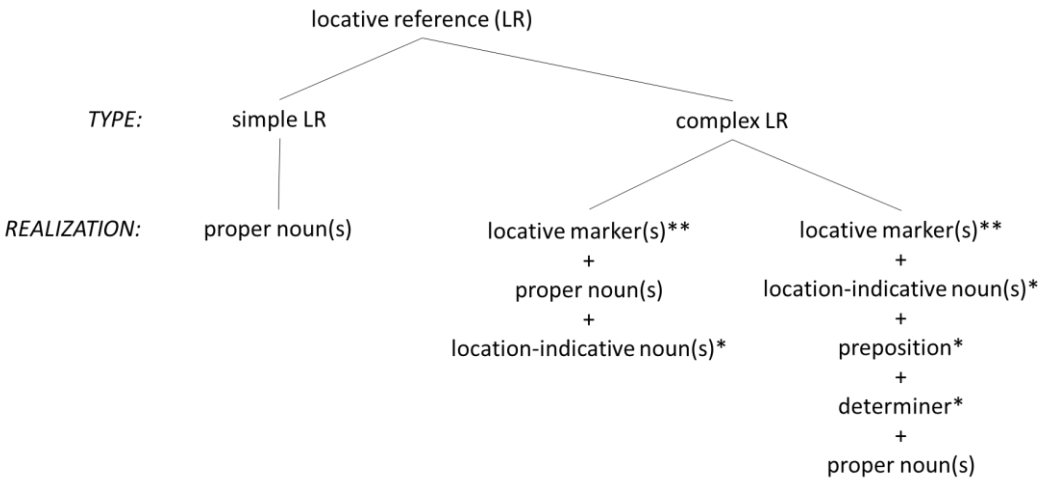

Figure 1: Phrasal structure of locative references

Examples of locative references are *Morocco*, *New York*, *south of Madrid*, *1km SW of Lake Henshaw*, *1h away from London*, *25min out of Melbourne*, *Mount Crawford*, *Bassmaya Project Power Plant station*, *province of Ontario*, *Jamia University*, *Dyckman Street Station*, *St Albans Park*, *Fox Valley Animal Referral Center*, *I 95 NB* and *George Washington bridge EB.*

In terms of semantics, a taxonomy was devised capturing the richness and variety of locative references, where abbreviations were also taken into account:

- Geopolitical entities: *country*, *state*, *region*, *province*, *city*, *town*, *kingdom*, *villa*, etc.
- Natural landforms: *mountain*, *mount*, *ridge*, *volcano*, *valley*, *lake*, *river*, *shore*, *beach*, *park*, *canyon*, etc.
- Points of Interest (POIs): *building*, *museum*, *school*, *station*, *stadium*, *garden*, *café*, *tavern*, *hospital*, *court*, *theater*, *residence*, *zoo*, *casino*, *square*, etc.

- Traffic ways (addresses, roads, highways): *street*, *st*, *boulevard*, *blvd*, *avenue*, *av*, *alley*, *road*, *rd*, *highway*, *hwy*, *freeway*, *fwy*, *turnpike*, *tpk*, *I(-)n*, *M(-)n* (where *n* represents a given number), etc.

Metonymic instances are a well-known issue in the literature when these represent the people of a place (e.g. *US officials*), organizations (e.g. *New Orleans Police Department*), government units (e.g. *London Councils*) or events (e.g. *New Zealand mass shooting*) (Liu *et al.* 2014; Gritta *et al.* 2018). They are borderline cases of locative references, and semantic boundaries are hard to establish (Wallgrün *et al.* 2018). A solution for this issue consisted in examining every ambiguous instance and determining, on the basis of the linguistic context, whether some degree of locative meaning was found in those references.

Given the importance of surrounding locative markers (e.g. *south of*, *northwest*, *25km away from*, *20 mins out of*, etc.), which contain more detailed information about the locational focus of a given incident, these were also annotated, following this taxonomy:

- Distance marker: *4 Kms from Narok Town*, *5miles from Dublin*, etc.
- Directional markers: *East Coast of Honshu*, *east of Exit 55*, *20 km NW of Durrës*, etc.
- Movement markers: *southbound I-91*, *northbound J19*, *eb J19*, etc.
- Temporal markers: *1h away from London*, *25min out of Melbourne*, etc.

3.1. Corpus compilation

The first stage involves decisions about text collection and corpus design. For collecting the tweets, which are the microtexts used in our corpus, the *FireAnt* tool was used (Anthony and Hardaker 2017) to obtain machine-readable tweet data, that is, JavaScript Object Annotation (JSON). The raw tweet data were collected on different dates after a keyword-based search using seven keywords related to crisis and emergency-related events which were *earthquake*, *flood*, *car accident*, *bombing attack*, *shooting attack*, *terrorist attack* and *incident*, so that tweets mentioning issues of different nature were extracted. The dates of extraction of the tweets were the following: 17 November 2019, 30 November 2019, 1 December 2019 and 2, 5 and 9 January 2020. In the corpus design substep, we also tackled key considerations such as what file formats were to be used and

what type of information would be included therein. The *FireAnt* tool provided not only the tweet text, but also the metadata associated with it. All those metadata were discarded and only tweet texts were saved in a .txt file before the preprocessing stage.

3.2. Corpus preprocessing

While the great majority of tweet texts contained one of the crisis-related keywords mentioned above, it was the case that some tweets were repeated on multiple occasions in retweets, split into different lines or empty. Our aim in this step was to obtain a representative corpus of unique tweets. The first preprocessing stage thus involved the following steps:

i) grouping multi-line tweets into a single line where each line represented one tweet by means of a regular expression that takes into account line breaks,

ii) removing retweets by means a regular expression that finds retweets and discards them, and

iii) removing duplicates and very similar tweets through a fuzzy matching algorithm (i.e. cosine distance similarity), which takes into account different combinations of characters and words in two strings to determine their degree of similarity.

Even though sensitive data about the tweets were removed by retaining the tweet text only, the text still contained sensitive information in the form of user mentions and URLs, which were dealt with in a second preprocessing stage. In this second preprocessing stage, non-standard linguistic features were kept in the tweets, too, since a key aspect in tweet location detection is to be able to overcome the challenge posed by non-standard uses of language. The main steps followed were the following:

i) Replacing user mentions and URLs by the tokens *user* and *URL*, respectively.
ii) Removing emojis and other special characters and leave punctuation marks and other commonly used characters (/, @, |…).
iii) Removing extra white spaces.
iv) Segmenting words contained in hashtags.

After both preprocessing stages, the resulting tweets were as unique as possible, clean and privacy-friendly. The released corpus contains the preprocessed tweets as such.

3.2. Corpus annotation

In the annotation stage, the corpus content was converted into a token-based tabular format with feature columns, separated by tabs, representing the following features: token, POS tag, presence in a *GeoNames*-based place-name dataset, presence in *WordNet*-based location-indicative noun dataset and part of a locative marker. In the last column, the class or label was tagged, following a Beginning-Medium-End-Single-Outside (BMESO) scheme, similar to others in Named Entity Recognition (NER), such as Beginning-Inside-Outside (BIO) (Jurafsky and Martin 2021). In other words, for multi-word locative references, the following labels were used: B_LOCATION, M_LOCATION and E_LOCATION. In the case of one-word locative references, S_LOCATION was used and, when a token or series of tokens are not locative references, O was used.

First, for preparing the annotated corpus, automatic tokenization and POS tagging were automatically applied, using the Stanford tokenizer and POS tagger functionalities (Toutanova and Manning 2000), and each tokenized tweet was delimited by a newline. The POS tags followed the *Penn Treebank* standard (Santorini 1990). Then, if tokens or a series of them were found in a place-name dataset obtained from *GeoNames* or in the location-indicative noun dataset obtained from *WordNet* (Vossen 1998) or were part of a locative marker, they were also automatically annotated with Boolean values: 0 if not present, 1 if present. This automatic tagging process was performed with the linguistic modules of LORE (Fernández-Martínez and Periñán-Pascual 2021a), which also, at last, detected and tagged the locative references found in the tweets. Table 1 presents an example of the token-based tabular format of the annotated corpus.

Token	**POS tag**	**Place-name dataset**	**Location-indicative noun dataset**	**Locative marker**	**Label**
Two	CD	0	0	0	O
vehicle	NN	0	0	0	O
incident	NN	0	0	0	O
,	,	0	0	0	O
48	CD	0	0	0	B_LOCATION
St	NNP	0	1	0	E_LOCATION
and	CC	0	0	0	O
32	CD	0	0	0	B_LOCATION
Ave	NN	1	1	0	M_LOCATION
NE	NNS	1	0	1	E_LOCATION

Table 1: Token-based tabular format in the FGLOCTweet Corpus

Since the accuracy of LORE is not perfect for detecting all and only locative references (precision score of 0.81 and recall score of 0.81 in Fernández-Martínez and Periñán-Pascual 2021a; precision score of 0.73 and recall score of 0.79 in Fernández-Martínez and Periñán-Pascual 2021b), the tags were manually revised for errors, such as missed locative references or wrongly assigned locative references, on the basis of the guidelines of the location types targeted by LORE. This was done using a common notepad editor tool (Notepad++). The POS tags were not revised since i) automatic POS tagging tools achieve a very high degree of accuracy (Manning 2011) and ii) feature noise is not a problem *per se* as long as other features can contribute in the learning process of the probabilistic-based model (Zhu and Wu 2004). POS tagging is a common component in NLP tasks and a typical feature, alongside tokenization, used in existing NER, since POS tags provide a strong linguistic cue for predicting the presence of named entities (Jurafsky and Martin 2021), especially because named entities are proper nouns. Particularly, in the case of locative references, these can be predicted by the presence of spatial prepositions (Hoang and Mothe 2018). However, automatic POS tagging might suffer from performance losses especially in the case of noisy text data (e.g. abbreviations, misspellings, ellipsis, etc.).

The other three features, presence in the place-name dataset obtained from *GeoNames*, presence in the location-indicative noun dataset retrieved from *WordNet* and being part of a locative marker, might also be prone to noise if a token or series of tokens are not found in these datasets. However, since different features are correlated, this noise might have a negligible impact in the training and evaluation phases of a location-detection system.[6]

Table 2 presents the distribution of locative references in terms of n-gram size in the corpus, whereas Table 3 provides statistical data about the number of locative references, number of tweets containing locative references, the average of locative references per tweet containing locative references and the average of locative references per tweet.

[6] In fact, it is known that in probabilistic-based models some degree of noise in a dataset can even be beneficial to avoid overfitting, that is, the memorization of the training dataset at the cost of performance degradation with new, unseen instances of data (Zur *et al.* 2009).

Number of unigrams (e.g. *Florida*)	3,256
Number of bigrams (e.g. *Grand Canyon*)	1,707
Number of trigrams (e.g. *St Albans Park*)	501
Number of n-grams where $n \geq 4$ (e.g. *Fox Valley Animal Referral Center*)	304
Total	**5,768**

Table 2: Distribution of locative references in terms of n-gram size in the corpus

Number of locative references	5,768
Number of tweets with locative references	3,416
Average of locative references per locative-rich tweet	1.69
Average of locative references per tweet	0.61

Table 3: Statistics about the number and average of locative references in the corpus

4. Discussion: Applications of the corpus, limitations and future research directions

The resulting corpus can then be split into two subcorpora: training and evaluation corpora using, roughly, an 85/15 split, which is the rule of thumb in the machine learning literature (Guyon 1997). The training corpus can be used to train a supervised probabilistic-based model for location detection, whereas the evaluation corpus can be used as a gold standard against which the output of a location-detection system can be tested for the evaluation of its accuracy (Pustejovsky and Stubbs 2013).

The main use of the FGLOCTweet Corpus is to build and assess location-detection models, either rule-based or probabilistic-based, for the task of identifying fine-grained locative references in crisis-related events from all over the world. Fine-grained detection of locative references is indeed a key aspect of accurate and useful location-detection systems which could potentially be used to save lives or prevent further damage to environmental or urban areas in crisis-related events by providing emergency responders with the location of a given incident. The corpus could also be used as a benchmarking dataset to compare the performance of different location-detection models, including named entity recognizers, too. Beyond that, linguists may find this corpus useful for approaching the conceptualization, expression and description of place in English during crisis-related events or even a general exploration of language use in microtexts dealing with crisis-related events.

In past research (Fernández-Martínez and Periñán-Pascual 2021a, 2021b), LORE, a rule-based model, and its probabilistic-based counterpart, neural LORE (nLORE), were built and assessed using this corpus, outperforming general-domain entity recognizers in benchmarking tests involving accuracy (i.e. precision and recall) and speed. A key

difference in the implementation of both models lies in how they make use of corpus data: the probabilistic-based model nLORE needed training data before the evaluation stage with the evaluation corpus, whereas the rule-based model LORE did not. This means that probabilistic-based models consume a lot of computational resources and time, whereas rule-based models tend to be much more efficient and quicker (Chiticariu *et al.* 2013). In this regard, LORE performed ten times faster than nLORE: it extracted locative references from around 7,000 tweets in a matter of 12 seconds as opposed to nLORE, for which it took almost two minutes. However, nLORE had slightly better accuracy than LORE in terms of precision (0.85 vs. 0.73), but lower recall (0.74 vs. 0.79).

As for the limitations, we would like to emphasize that, even though the Stanford POS tagger may achieve a very high accuracy of 97 percent (Manning 2011), its accuracy might have been somewhat lower with the tweets, introducing some corpus noise in the POS tags feature. A rule-based model that is assessed on our corpus, if developed with lexicogrammatical rules taking into account grammatical categories, might be misled by wrong POS tags and extract wrong items or miss potential locative references, if it does not rely on the other corpus features as well. The probabilistic-based model would not, however, be impeded by this providing that it makes use of the different features at the same time, and even in that case, some degree of corpus noise, as stated above, might be beneficial to avoid overfitting in the training phase of the probabilistic-based model.

Further research lines could be pursued, especially in a time where novel NLP approaches employing transformers like BERT (Devlin *et al.* 2018) show promising results, which could be fine-tuned using our corpus. Also, LORE could be employed to automatically aggregate new annotated data to the corpus in an unsupervised fashion, thus enriching the number and variety of locative references, though at the cost of introducing corpus noise. In such a scenario, it can be insightful to analyze whether corpus noise in larger sizes of annotated corpus data could be detrimental to the performance of a model trained on and assessed with these data. Also, transfer learning techniques together with this unsupervised aggregate of new data could be used to create a multilingual corpus for multilingual fine-grained location-detection tasks, including and mixing, for instance, annotated Spanish and French tweets.

5. Conclusion

Location detection in social media is still an unsolved task in NLP. Since there is a growing need to automatically obtain actionable information from social media in emergency-based contexts where granularity and time play an essential role to understand the *where* of an incident, the development of fine-grained annotated corpus data becomes of utmost importance, especially to train and assess location-detection models. Despite that, available corpora are lacking up to this date; annotation standards are different, and many location types are, at best, poorly addressed in the literature or, at worst, neglected. Besides, phrasal structures indicating distance (e.g. *n kms away from X*), direction (e.g. *south of X*), movement (e.g. *eastbound*) and time (e.g. *n mins out of X*), which may take part in locative references, are not annotated, thus missing very detailed geospatial information which could be highly important in crisis-related events. To address this gap, the FGLOCTweet Corpus is presented, with the aim of providing an English tweet-based corpus for fine-grained location-detection tasks to advance research in the NLP and linguistic communities.

References

Ahlers, Dirk. 2013. Assessment of the accuracy of GeoNames gazetteer data. In Chris Jones and Ross Purves eds. *Proceedings of the 7th Workshop on Geographic Information Retrieval - GIR '13*. New York: Association for Computing Machinery, 74–81.

Ahmed, Mohammed F., Lelitha Vanajakshi and Ramasubramanian Suriyanarayanan. 2019. Real-time traffic congestion information from tweets using supervised and unsupervised machine learning techniques. *Transportation in Developing Economies* 5/2: Article 20. https://link.springer.com/article/10.1007/s40890-019-0088-2 (10 September, 2021.)

Anthony, Laurence and Claire Hardaker. 2017. *FireAnt* (Version 1.1.4). Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software (10 September, 2021.)

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In Ruslan Mitkov and Jong C. Park eds. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, 356–364. http://www.aclweb.org/anthology/I13-1041 (10 September, 2021.)

Chiticariu, Laura, Yunyao Li and Frederick R. Reiss. 2013. Rule-based information extraction is dead! Long live rule-based information extraction systems! In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu and Steven Bethard eds. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. New York: Association for Computational Linguistics, 827–

832.
Crooks, Andrew, Arie Croitoru, Anthony Stefanidis and Jacek Radzikowski. 2013. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17/1: 124–147.
Das, Raul D. and Ross S. Purves. 2019. Exploring the potential of Twitter to understand traffic events and their locations in greater Mumbai, India. *IEEE Transactions on Intelligent Transportation Systems* 21/12: 1–10.
De Bruijn, Jens A., Hans de Moel, Brenden Jongman, Jurgen Wagemaker and Jeroen C. Aerts. 2018. TAGGS: Grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis* 2/2: 1–14.
Developer Policy – Twitter Developers. 2021. Twitter developer platform. https://developer.twitter.com/en/developer-terms/policy (5 December, 2021.)
Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv* http://arxiv.org/abs/1810.04805 (10 September, 2021.)
Dredze, Mark, Michael J. Paul, Shane Bergsma and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In Martin Michalowski, Wojtek Michalowski, Dympna O'Sullivan, Szymon Wilk eds. *Expanding the Boundaries of Health Informatics Using Artificial Intelligence: Papers from the Association for the Advancement of Artificial Intelligence 2013 Workshop*. Palo Alto, California: Association for the Advancement of Artificial Intelligence, 20–24. https://www.aaai.org/ocs/index.php/WS/AAAIW13/paper/view/7085
Eisenstein, Jacob. 2013. What to do about bad language on the internet. In Lucy Vanderwende, Hal Daumé III and Katrin Kirchhoff eds. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New York: Association for Computational Linguistics, 359–369. https://aclanthology.org/N13-1037/
Eke, Paul I. 2011. Using social media for research and public health surveillance. *Journal of Dental Research* 90/9: 1045–1046.
Fernández-Martínez, Nicolás José and Carlos Periñán-Pascual. 2021a. LORE: A model for the detection of fine-grained locative references in tweets. *Onomazein* 52: 195–225.
Fernández-Martínez, Nicolás José and Carlos Periñán-Pascual. 2021b. nLORE: A linguistically rich deep-learning system for locative-reference extraction in tweets. In Engie Bashir and Mitja Luštrek eds. *Intelligent Environments 2021: Workshop Proceedings of the 1st International Workshop on Artificial Intelligence and Machine Learning for Emerging Topics (ALLEGET '21)*. Amsterdam: IOS Press 243–254.
Gonzalez-Paule, Jorge David, Yeran Sun and Yashar Moshfeghi. 2019. On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing and Management* 56/3: 1–14.
Gritta, Milan, Moahammad T. Pilehvar, Nut Limsopatham and Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation* 52/2: 603–623.
Guyon, Isabelle. 1997. A scaling law for the validation-set training-set size ratio. Technical report. Berkeley, California: *AT&T Bell Laboratories* 1–11. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1337&rep=rep1&type=pdf (10 September, 2021.)
Hoang, Thi B. N. and Josiane Mothe. 2018. Location extraction from tweets. *Information Processing and Management* 54/2: 129–144.

Hu, Yingjie and Jimin Wang. 2020. How do people describe locations during a natural disaster: An analysis of tweets from Hurricane Harvey. In Krzysztof Janowicz and Judith A. Verstegen eds. *11th International Conference on Geographic Information Science (GIScience 2021)*. Dagstuhl, Germany: Dagstuhl Publishing Company, 6.1–6.16. https://drops.dagstuhl.de/opus/volltexte/2020/13041/pdf/LIPIcs-GIScience-2021-I-6.pdf (10 September, 2021.)

Imran, Muhammad, Carlos Castillo, Fernando Diaz and Sarah Vieweg. 2014. Processing social media messages in mass emergency: Survey summary. *WWW'18: Companion Proceedings of the The Web Conference 2018*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 507–511. https://dl.acm.org/doi/10.1145/3184558.3186242 (10 September, 2021.)

Inkpen, Diana, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi and Diman Ghazi. 2017. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems* 49/2: 237–253.

Jongman, Brenden, Jurgen Wagemaker, Beatriz Romero and Erin de Perez. 2015. Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and Twitter signals. *ISPRS International Journal of Geo-Information* 4/4: 2246–2266.

Jurafsky, Daniel and James H. Martin. 2021. Sequence labeling for parts of speech and named entities. In Dan Jurafsky and Hames H. Martin eds. *Speech and Language Processing*: 1–27. https://web.stanford.edu/~jurafsky/slp3/8.pdf (10 September, 2021.)

Khodabandeh-Shahraki, Zahra, Afsaneh Fatemi and Hadi Tabatabaee-Malazi. 2019. Evidential fine-grained event localization using Twitter. *Information Processing and Management* 56/6: Article 102045.

Liu, Fei, Maria Vasardani and Timothy Baldwin. 2014. Automatic identification of locative expressions from social media text. In Dirk Ahlers ed. *LocWeb '14: Proceedings of the 4th International Workshop on Location and the Web*. New York: Association for Computing Machinery, 9–16.

Manning, Christopher D. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh ed. *Computational Linguistics and Intelligent Text Processing.* Berlin: Springer Berlin Heidelberg, 171–189.

Martínez-Rojas, María, María del Carmen Pardo-Ferreira and Juan Carlos Rubio-Romero. 2018. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management* 43: 196–208.

Middleton, Stuart E., Giorgos Kordopatis-Zilos, Symeon Papadopoulos and Yiannis Kompatsiaris. 2018. Location Extraction from Social Media. *ACM Transactions on Information Systems* 36/4: 1–27.

Mourad, Ahmed, Falk Scholer, Walid Magdy and Mark Sanderson. 2019. A practical guide for the effective evaluation of Twitter user geolocation. *ACM Transactions on Social Computing* 2/3: 1–23.

Purves, Ross S., Paul Clough, Christopher B. Jones, Mark H. Hall and Vanessa Murdock. 2018. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval* 12/2–3: 164–318.

Pustejovsky, James and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning: A Guide to Corpus-building for Applications*. Sebastopol, California: O'Reilly Media, Inc.

Rayson, Paul. 2014. Computational tools and methods for corpus compilation and

analysis. In Douglas Biber and Randi Reppen eds. *The Cambridge Handbook of English Corpus Linguistic*. Cambridge: Cambridge University Press, 32–50.

Reppen, Randi. 2010. Building a corpus. In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 31–37.

Santorini, Beatrice. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* 3rd revision, 2nd printing. Department of Computer and Information Science, University of Pennsylvania: Technical Report MS-CIS-9047. https://repository.upenn.edu/cis_reports/570/ (10 September, 2021.)

Singh, Lisa, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter [preprint 31 March 2020]. *ArXiv*. http://arxiv.org/abs/2003.13907 (10 September, 2021.)

Siriaraya, Panote, Yihong Zhang, Yuanyuan Wang, Yukiko Kawai, Mohit Mittal, Péter Jeszenszky and Adam Jatowt. 2019. Witnessing crime through tweets. In Farnoush Banaei-Kashani, Goce Trajcevski, Ralf Hartmut Güting, Lars Kulik and Shawn Newsam eds. *SIGSPATIAL '19: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York: Association for Computing Machinery, 568–571.

Toutanova, Kristina and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Hinrich Schütze and Keh-Yih Su eds. *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. New York: Association for Computational Linguistics, 63–70.

Vieweg, Sarah, Amanda L. Hughes, Kate Starbird and Leysia Palen. 2010. Microblogging during two natural hazards events. In Elizabeth Mynatt ed. *CHI '10 Proceedings of the 28th International Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 1079–1088.

Vossen, Piek. 1998. Introduction to EuroWordNet. *Computers and the Humanities* 32/2–3: 73–89.

Wallgrün, Jan Oliver, Morteza Karimzadeh, Alan M. MacEachren and Scott Pezanowski. 2018. GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science* 32/1: 1–29.

Wang, Jimin and Yingjie Hu. 2019. Are we there yet ? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In Bruno Martins ed. *GeoHumanities '19 Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*. New York: Association for Computing Machinery, Article 2, 1–6.

Zhang, Cheng, Chao Fan, Wenlin Yao, Xia Hu and Ali Mostafavi. 2019. Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management* 49: 190–207.

Zhu, Xingquan and Xindong Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review* 22/3: 177–210.

Zinsmeister, Heike, Erhard Hinrichs, Sandra Kübler and Andreas Witt. 2009. Linguistically annotated corpora: Quality assurance, reusability and sustainability. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook Vol. 1*. Berlin: Walter de Gruyter, 759–772.

Zur, Richard M., Yulei Jiang, Lorenzo L. Pesce and Karen Drukker. 2009. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics* 36/10: 4810–4818.

Corresponding author
Nicolás José Fernández-Martínez
Department of Languages
Campus de los Jerónimos, Guadalupe 30107
Murcia
e-mail: njfernandez@ucam.edu

received: October 2021
accepted: December 2021