Automated construction of a French Entity Linking dataset to geolocate social network posts in the context of natural disasters

Gaëtan Caillaut, Cécile Gracianne, Nathalie Abadie, Guillaume Touya, Samuel Auclair

Automated construction of a French Entity Linking dataset to geolocate social network posts in the context of natural disasters

Gaëtan Caillaut*
BRGM
g.caillaut@brgm.fr

Cécile Gracianne
BRGM
c.gracianne@brgm.fr

Nathalie Abadie
LASTIG, Univ Gustave Eiffel, IGN-ENSG
nathalie-f.abadie@ign.fr

Guillaume Touya
LASTIG, Univ Gustave Eiffel, IGN-ENSG
guillaume.touya@ign.fr

Samuel Auclair
BRGM
s.auclair@brgm.fr

ABSTRACT

During natural disasters, automatic information extraction from Twitter posts is a valuable way to get a better overview of the field situation. This information has to be geolocated to support effective actions, but for the vast majority of tweets, spatial information has to be extracted from texts content. Despite the remarkable advances of the Natural Language Processing field, this task is still challenging for current state-of-the-art models because they are not necessarily trained on Twitter data and because high quality annotated data are still lacking for low resources languages. This research in progress address this gap describing an analytic pipeline able to automatically extract geolocatable entities from texts and to annotate them by aligning them with the entities present in Wikipedia/Wikidata resources. We present a new dataset for Entity Linking on French texts as preliminary results, and discuss research perspectives for enhancements over current state-of-the-art modeling for this task.

Keywords

Automated geotagging, French Entity Linking, Wikipedia, Twitter, Crisis Management, Natural Disaster

INTRODUCTION

When a natural disaster occurs, the main issue is to make sense of what is going on, with the necessity to be able to qualify the severity of the situation and to follow its evolution over time. Response organizations address this issue by collecting as much information as possible in order to increase their "situational awareness" and to build a realistic "common operational picture" on which to undertake effective actions. After having considered for a long time that the only exploitable data for disaster management should come from specialized services guaranteeing their reliability and their technical quality, the experience of recent years has shown that citizen data could allow - when well exploited - to bring additional knowledge and to build better targeted response (Kaufhold 2021). Federating communities of users accustomed to daily comment events on live, social media channel the data of millions of connected "citizen sensors" (Goodchild 2007) endowed with five senses, able to share testimonies spontaneously and quickly. In practice, the richest information usually comes from those citizens being closest to the area affected

*corresponding author

by the disaster. Because the natural disaster affects their immediate environment, these "local citizens" (Grace et al. 2017) are indeed more inclined - both in the physical and digital spheres - to help or to exchange objective information on the field situation (Akter and Wamba 2019; Starbird et al. 2012). The latter tend to exchange information on the effects of disasters, victims or other types of useful information, while populations farther away from events relay this information, or express their empathy to the victims (Olteanu et al. 2015). For these reasons, monitoring and analysis of social media has become at the heart of the concerns of crisis managers (Rasmussen and Ihlen 2017), especially Twitter. Although it is not the most widely used social media, the Twitter platform present particularly useful features for reporting and monitoring natural disasters such as publication of short messages in real time, free streaming Application Programming Interface (API) making it possible to automate monitoring tasks (Auclair et al. 2019), and ability to attach pictures. Its real-time monitoring and analysis may turns in delineating rapidly the extension of the area impacted by an earthquake (Fayjaloun et al. 2021) or a flood (Arthur et al. 2018), or identifying emergency situations of interest for the rescue or security forces (Qadir et al. 2016; Z. Wang and Ye 2018).

As such, several works considering information extraction for natural disaster crisis management tried to place their results on a map to make it directly usable (Stanek and Drosio 2012; Zhang et al. 2016) or advocate for alternative way to build those maps such as crowdsourcing mapping (Hunt and Specht 2019). A very low proportion of tweets have an intrinsic geolocation (less than 1 %), like GPS coordinates (Cheng et al. 2010). Consequently, geolocation of the information should be inferred from the information present in the text and/or the metadata of the tweet itself. This can be done thanks to the recent advances in Natural Language Processing (NLP) allowing to extract accurate word representations, in a vectorial form, from raw text. These vectors, or *word embeddings*, hold precious semantic and syntactic latent features, including geospatial features. While the first word embedding approaches resulted in a fixed set of vector for a given vocabulary (Mikolov et al. 2013; Pennington et al. 2014), modern approaches (Devlin et al. 2019) propose to first train a generic model and then to fine-tune it on any downstream task. Among others, works have been carried out to fine-tune such systems to solve the well-known Named Entity Recognition (NER) and the Entity Linking (EL) tasks. The first one aims to detect and classify named entities, such as persons or organizations, in running text, while the second one aims to disambiguate these entities by linking them to an identifier, usually a node in a Knowledge Base (KB). This allows enriching the text with external information which can help either an automated system to provide a better prediction or a human agent to better understand, or to refute, the outputs of such systems.

In this work, we propose to apply EL on toponyms, or any geolocatable entity, mentionned in tweets written during crisis such as earthquakes or floods. We use gazeteers, such as OpenStreetMap, as target for our EL system. Such a system should enable end-users to collect fine to coarse grained information about the spatial entities mentioned in a tweet, such as, most importantly, its coordinates on Earth, but also its area or its population. Depending on these information, rescue teams may behave differently. For instance, an intervention in a densely populated city center does not require the same human and material resources than another one in a less populated residential neighborhood. However, training such a system requires quite a large annotated corpus, which makes solving this task a real challenge, especially when dealing with French tweets. Indeed, French resources tend to be less numerous than their English counterparts. This is why we first tackle the problem of collecting and annotating a French corpus dedicated to the Entity Linking (EL) task. Due to the scarcity of geotagged tweets, collection and annotation must be made manually, which is heavily resources consuming. Following a long line of works (Logan et al. 2019; Merity et al. 2016; Bunescu and Paşca 2006), we propose an automated alternative using hyperlinks in Wikipedia pages as a supervision signal.

Such a dataset would probably not fit Twitter data, therefore we plan, in a second time, to augment our dataset with Twitter data. Nevertheless, it has already been shown that deep neural networks can be effectively pre-trained on a generic task then fine-tuned to fit other, more specific, tasks, the most obvious example being `BERT`. Rajapaksha et al. (2021) successfully trained large Transformer models to detect clickbait on Twitter data even though the base pre-trained models were not necessarily trained on Twitter. Also, J. Wang et al. (2020) showed that augmenting a Twitter dataset with data from Wikipedia had a positive impact on the predictive performance of their system. We will first do a quick review of works proposing solution to extract spatial features from raw texts. Then, we will show the benefits of an EL system over the others and we will describe the process we used to automatically build our annotated dataset for EL. Finally, we will introduce our system architecture, which should fix some limitation of current state-of-the-art approaches.

RELATED WORKS

Since the recent advances in Natural Language Processing (NLP), Named Entity Recognition approaches have been mostly relying on deep learning architectures (Li et al. 2020). Modern neural architectures produce contextual

representations of words, that can be leveraged to classify token and identify those representing spatial named entities, yielding state-of-the-art results. The most famous approaches are `BERT` (Devlin et al. 2019) and `ELMO` (Peters et al. 2018), which rely, respectively, on the Transformer (Vaswani et al. 2017) and LSTM architectures (Gers et al. 2000). Spatial Named Entity Recognition is a popular subtask of NER that can be performed with the same neural architectures but requires dedicated annotated corpora for training and testing. J. Wang et al. (2020) relied on a subset of Wikipedia to generate automatically a dataset to train a toponym recognition model. They trained a model on this dataset, another on the `WNUT2017` Twitter Dataset (Derczynski et al. 2017), and a last one on both. They evaluated their models on a Twitter dataset extracted during Hurricane Harvey, in 2017, and showed that the model trained on Wikipedia performs worse than the model trained on `WNUT2017`. However, using data from Wikipedia is far from useless since the models trained on both corpora surpass them by a large margin.

These architectures can be further specialized and fine-tuned for geotagging purposes. This means associating the tokens representing a spatial named entity, no longer with a class label, but either with predicted geographical coordinates or with the identifier of a geographical entity of known location listed in a gazetteer. In both cases, leveraging the contextual distributed representations of toponyms produced by modern language models may prove especially useful. Indeed, toponyms often refer to multiple places in the world, for instance, five French municipalities are called *Chaumont*. Previous works in the field of spatial entity resolution, either based on supervised approaches or not, have shown the benefits of using contextual knowledge (nearest neighbors, spatial relations, popularity, importance, etc. . .) for place names disambiguation. `BERT` or `ELMO` based language models have been pre-trained and made available for many languages. Geotagging systems can hence be derived from such models.

Approaches for geographic coordinates prediction or geographic region classification. The quality of the results of gazetteer-based approaches depends strongly on the quality of the input gazetteer, and in particular on its completeness. To overcome this difficulty, especially for areas where few location data are available, many text geotagging approaches propose to first discretize the surface of the Earth, or the studied area, into cells. In this context, the geotagging task is a classification task with n classes, n being the number of cells. Early approaches of that kind, are presented in the survey proposed by Melo and Martins (2017). More recently, deep learning architectures have been proposed to solve this classification task, like in Gritta et al. (2018) or in Yan et al. (2021). In both approaches, some geographical knowledge is introduced by feeding the model with the same spatial grid where each cell is filled with knowledge regarding the candidate geographic entities: in the former, the candidates population count is added, in the later it is the frequency of co-occuring place mentions. None of these approaches can predict an exact location, since their theoretical precision is bound to the cell size, even though it is often the coordinates of the cell's barycenter that are given as results. This can be somehow mitigated by modulating the cell size according to some criteria: for instance, it seems reasonable to set a smaller cell size in an area densely populated. This kind of approach can also be used to predict a vague location that can then be leveraged by another, more precise, geotagging method. Such an approach is proposed by Cardoso et al. (2022). A first deep learning architecture is used to predict a probability distribution over geo-spatial regions in a hierarchical spatial grid and this result is then combined with the centroid coordinates of the grid cells and a second loss function to predict geographical coordinates.

Gazetteer-based approaches. The last kind of geotagging approaches takes advantage of gazetteers, such as Open Street Map[1] (OSM), Geonames[2] or BDTOPO ®[3]. Gazetteers store the names of any type of geographic entity, such as cities, streets, lakes or buildings, with at least a spatial reference alongside (*i.e.*: coordinates, a geometry, etc. . .). Further information regarding geographic entity properties may be added, like their nature, their geometric properties, their population, etc. Such information, providing some contextual knowledge about geographic entities, may reveal useful for place name disambiguation. That is why many works have focused on gazetteer construction to improve spatial named entity linking results (Overell and Rüger 2008; Brando et al. 2015; Spitz et al. 2016; Kim et al. 2017; Ardanuy and Sporleder 2017). Besides, gazetteers may provide very accurate location information for each place name, which can be crucial for emergency response applications. Predicting such geographic entities is the same as solving Entity Linking (EL), a task consisting in aligning text and entities from knowledge bases. Most of the previous works have focused on unsupervised approaches based on two steps :

- Candidates selection: For each spatial named entity mention, this step aims at selecting the most similar place names in the gazetteer. It is mostly performed with string similarity measures (Recchia and Louwerse 2013).
- Candidates ranking: This step aims at finding, among the selected gazetteer entries, the most likely to be represented the same place as the spatial named entity mentioned in the text. This is performed either by

[1] https://www.openstreetmap.org/
[2] http://www.geonames.org/
[3] https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo

applying a set of heuristics based on geographic entity distances (Habib and Keulen 2012; Derungs and Purves 2014; Blank and Henrich 2015), their spatial relations (Kim et al. 2017; Paris et al. 2017), their importance (Amitay et al. 2004; Brando et al. 2015) or with approaches implementing supervised learning algorithms, like SVM or LambdaMART, with hand-crafted features to describe spatial named entity mentions and geographic entities listed in the gazetteer (Martins et al. 2010; Daiber et al. 2013; Speriosu and Baldridge 2013; Santos et al. 2015). Many works use a corpus built from the English Wikipedia to train their learning algorithm.

More recently, some approaches have tried to solve this task with deep-learning-based classification methods. Xu et al. (2019) proposes DLocRL, a deep learning pipeline for locations recognition and linking in tweets, that computes a matching score between a given tweet and a location profile built from geographic data extracted from Foursquare[4]. They also add a post-processing step based on a Geographical Pair Linking algorithm that leverages the geographical coherence between co-occurring spatial entity mentions. To overcome the limitation due to the overwhelming number of classes, some works propose to train two classifiers — the first one receiving the text containing (spatial) entities and the second one the features of the (spatial) entities to predict — in parallel to predict the same spatial embedding (Botha et al. 2020).

Whatever the strategy adopted, there are two major challenges for geotagging applications based on deep learning: integrating geographical knowledge into the model and building an appropriate corpus to train the model.

USING WIKIPEDIA TO TRAIN A SPATIAL ENTITY LINKING PIPELINE

As suggested in the previous section, predicting entities from KB or gazeteers, such as OSM, would have several benefits. Amongst them, gazeteers store many precious information that could be extremely valuable in case of an emergency, such as, for instance, the height of a building or the presence of a gas pipe near in a damaged area. Unfortunately, there are very few datasets mapping text spans to OSM entities, making it impossible to train an EL system to predict OSM entities. Actually, to the best of our knowledge, such mappings exist only between the couple Wikipedia/Wikidata. For that reason, Wikipedia and Wikidata are ideal resources to train EL systems, since all Wikipedia pages have a corresponding entity in the Wikidata graph. One can use the internal hyperlinks in Wikipedia pages to train a system to (1) detect entities, since they are identified by hyperlinks, and (2) align them to the Wikidata graph. Actually, many works have been relying on Wikipedia to extract supervisions signals from hyperlinks (Botha et al. 2020; Merity et al. 2016; Ghaddar and Langlais 2017; Nothman et al. 2013), but only a few uses the French Wikipedia. Other works have been using Wikipedia to train models capable of resolving, specifically, toponyms and others geolocatable entities (Martins et al. 2010; Geiß et al. 2015). In the case of a spatial entity, the Wikidata graph often contains links to other geospatial databases, such as GeoNames or OSM, making it possible to retrieve fine-grained spatial features. Even if the Wikidata entity lack these links, spatial features can, most of the time, be retrieved directly from the Wikidata KB.

However, models trained on Wikipedia may not be well-suited to deal with social network data, since Wikipedia pages and social network posts generally do not share the same writing style. It is still possible to alter the Wikipedia pages, for instance by randomly removing or swapping characters, or changing cases, which should make the model more robust. But such alterations could conflict with the model's tokenization methods, which could results in a lot of <unk> token (J. Wang et al. 2020), which does not hold any useful information. Furthermore, such artificial data augmentation are not sufficient since they cannot generate data with the same vocabulary and slang used on social networks. That being said, the lack of annotated social network corpora, especially in French, will inevitably force us to rely on the Wikipedia dataset, since it is the only publicly available corpus with direct mappings between entities and text spans. Nevertheless, Wikipedia is insufficient on its own. As pointed by J. Wang et al. (2020), the best performances are obtained with a combination of the Wikipedia dataset and a small social network dataset, `WNUT2017` (Derczynski et al. 2017).

Building an Entity Linking dataset automaticaly

The Wikimedia foundation provides dumps of Wikipedia[5] and Wikidata[6]. The Wikidata dump consists in a big JSON file whose size is approximately 70 GB as of November 2021. It is relatively easy to read since it is encoded as a single JSON array and entities can be processed incrementally by reading the file line by line. Hence, to process the file, one needs to read the dump line by line and parse the JSON string to get a representation of the current

[4] https://foursquare.com/
[5] https://dumps.wikimedia.org/
[6] https://www.wikidata.org/wiki/Wikidata:Database_download

Wikidata entry. The sole difficulties lie in understanding the Wikidata architecture, which is not really user-friendly, and in processing efficiently and quickly such a large dataset.

The Wikipedia dumps are released as separate XML files: one per language. The French Wikipedia dump, as of November 2021, is around 5.1 GB. Being plain, standard, XML, the dump is quite easy to read too, although, as we will see in the following, it is not really usable. Indeed, Wikipedia pages are written in a specific language, the *wiki markup* language, also known as *wikitext* or *wikicode*. This language is known to have no clearly defined syntax and extensively rely on (nested) templates expansion, making it extremely difficult to write a reliable parser[7]. Indeed, previous work shows that standard parsing methods could not reproduce the expected Wikipedia output (Dohrn and Riehle 2011). As a consequence, researchers seeking to leverage the Wikipedia corpora only can use unreliable tools. Currently, `WikiExtractor` (Attardi 2015) seems to be the most used software to extract text from Wikipedia dumps, despite that a lot of texts were removed because the tool could not handle some templates from the French Wikipedia dump. `Wikitextprocessor`[8] seems to be a very promising tool since it is capable of expanding templates much more reliably, at the cost of an non-negligible increase in processing time. However it does not yet handle non-English Wikipedia dumps, so it is not suitable for applications on other languages.

We also explored solutions using the Wikipedia API. Indeed, it is possible to retrieve the wikitext of a page using the API, then ask the API to do template expansion on the wikitext. However, the template expansion API is limiting the size of its input, and rate limits disallow making too much queries. These limitations could be bypassed by hosting our own Wikipedia instance. While running a Wiki instance is made easy by using the official Docker image, importing the whole Wikipedia dump is not feasible in a reasonable time frame since development of the import tool have been abandoned[9].

In the end, we did not have much choices but to rely on the HTML output of the Wikipedia website. However, scrapping the whole website is excluded since it would take too much time, and it would be inappropriate regarding the number of queries made to the Wikipedia web server. Furthermore, we do not need the full Wikipedia corpus since we aim working with social network data. Hence, training a model on the entire Wikipedia corpus may introduce biases into the system, making it inefficient on social network data. So, we extracted a small subset of the French Wikipedia corpus following the same scheme as Merity et al. (2016). Using `Pywikibot`, we extracted the list of *featured* and *good* articles, which are articles tagged as being well-written by real users from the Wikipedia community. This resulted in 6027 articles, as of January 2022. We then scanned these pages in order to extract links to other Wikipedia pages. These links are very important, since we will use them as gold reference to train our Entity Linking system. As we intend to use a system similar to the one proposed by Botha et al. (2020), we also need a short description per entities. So, we also downloaded the pages referring to the entities we extracted from the 6027 *featured* and *good* articles, resulting in 336 743 articles, from which we extracted the first paragraph since it is supposed to be a short description of the page, according to the Wikipedia's guidelines[10].

Not so much cleaning was necessary to make the HTML pages usable, since most of the work have been done by the Wikipedia backend. We removed everything that is not raw text, such as tables and pictures and replaced the mathematical formula by the textual representation, provided by Wikipedia, intended to text-based web browser. We also removed common sections, such as "references" or "see also" sections, which we think are irrelevant for most of the use-cases. Finally, we tagged the links with `[E]` tags. For instance, if a link to the page `Paris` with the text "Ville lumière" (City of light) is found, it is replaced by the string "[E=Paris]Ville lumière[/E]". This allow to preserve both the original text, as well as the title of the linked page.

For each page we extracted from Wikipedia, we also extracted the following properties from the linked Wikidata entity:

QID A string identifying a Wikidata entity. For instance, the QID of `Paris` is `Q90`.

Description A short description of the entity. Not always reliable.

Label The name of the entity. For instance the label of the entity `Paris` is "Paris".

Aliases A list of aliases for the entity. For instance, both "Paname" or "Ville-Lumière" refer to the entity `Paris`.

Type A *suggested* type for the entity. One of "GEOLOC", "ORG", "PERSON", "DATE" or "OTHER".

[7] https://utcc.utoronto.ca/~cks/space/blog/programming/ParsingWikitext
[8] https://github.com/tatuylonen/wikitextprocessor
[9] https://www.mediawiki.org/wiki/Manual:MWDumper
[10] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

Entity Type	Count	Instances
PERSON	85 670	301 714
GEOLOC	98 749	552 583
ORG	2180	13 849
DATE	2024	38 816
OTHER	116 118	712 683

Table 1. Number of distinct entities' types as well as their appearance in the corpus.

The *Type* property is not extracted directly from Wikidata. We built an *ad hoc* set of rules to predict the type of an entity. These rules check the presence of some properties in the entities. For instance, the entity `Paris` is an instance of (property `P31`) of the entity `Capital` (whose QID is `Q5119`). Then, it is probably a spatial entity, so we set the value "GEOLOC" to the *Type* property. The number of instances of each entity type is shown in Table 1, but the data extracted from our set of rules has not yet been audited, so we do not recommend relying on it for a production system. Our classification probably needs further refinement considering the high number of "OTHER" entities.

We released our dataset on the HuggingFace hub[11]. It contains **6023** documents and **304 826** distinct entities. Among the **37 051 326** words, **2 913 959** are part of an entity (an entity can be made of multiples words). A total of **1 619 961** entities have been annotated. The scripts we used to produce this dataset are available as a Git repository[12]. Currently, the scripts are specific to the French Wikipedia, since some links are hard-coded, as well as the name of the sections to remove during the cleaning phase. However, it should not be too difficult to edit the scripts in order to make them work with any language.

A PIPELINE FOR NAMED ENTITY RECOGNITION AND ENTITY LINKING

We build a system heavily inspired from the work of Botha et al. (2020). This system encompass two Transformer-base encoder trained conjointly to produce similar outputs. The first encoder is called the *mention encoder* and the second one the *entity encoder*. In the original work, the mention encoder takes as inputs the title of a Wikipedia page and a sentence from the same page, in which **one** entity is tagged. The entity encoder takes as input a textual description of the tagged entity. Their system achieve impressive results, especially considering that it was able to produce accurate representations for many languages, even on languages on which it has not seen during training. However, we argue that this system, as is, is not well-suited for real-world applications. First, the system can only process sentences from Wikipedia, since the Wikipedia title is a part of its inputs. As such, it may be difficult to use this system on Twitter data. Second, the system can compute a representation for **a single entity** that has to be **tagged beforehand**. Hence, this system has to be used in conjunction with an accurate Named Entity Recognition (NER) system. Yet, we think that doing both NER and EL in the same time may be preferable because it may reduce both the size of the model and processing time. Indeed, the dual encoder architecture is already heavy, since it encompass two distinct `BERT` models[13]. Adding a third system to do NER would increase significantly the size and complexity of an already complex architecture.

To alleviate the two mentioned limitations, we first propose to not rely the mention encoder on Wikipedia specific features, such as the title of a page. In this case, the mention encoder takes only text as input. Second, the mention encoder should not be limited to only one entity per input sentence and be designed to (i) detect any Named Entities and (ii) compute Entity embeddings for each of them. In this configuration, the system has to be trained to minimize the following dual objective:

$$NER(o_{ner}, y_{ner}) + EL(m_{el}, e_{el}) \tag{1}$$

Where o_{ner} are the predicted NER labels, y_{ner} are the expected NER labels, m_{el} are the output entity representations from the mention encoder and e_{el} are the output entity representation from the entity encoder. The NER function is the classical cross-entropy loss and EL is a loss function based on the cosine similarity.

CONCLUSION

Geotagging social network posts requires to be able to extract accurate and relevant spatial features from raw text. Several strategy exists, and we chose to train entity embeddings by leveraging links found in Wikipedia pages. In

[11] https://huggingface.co/datasets/gcaillaut/frwiki_good_pages_el
[12] https://github.com/GaaH/frwiki_good_pages_el
[13] We suppose this is why the authors use only the four first `BERT`'s layers

this work, we propose a set of open-source tools to automatically scrap data from Wikipedia and annotate them using links found in Wikipedia articles. We also release a French dataset from the *featured* and *good* articles with both NER and EL annotations.

While we do not have, yet, experimental results, we propose nevertheless further developments of the method proposed by Botha et al. (2020) to make it more suitable in real-world applications since it does not rely specifically on Wikipedia metadata (such as the page's title) nor requires entity annotations. Experiments need to be carried out to validate our proposed pipeline.

Future works will be dedicated to extending our dataset. First, we will manually annotate a set of French tweets extracted during natural disaster events, such as the Alex storm that hits France in September 2020. Such efforts are critical, since, as stated previously, a good system trained only on Wikipedia (or any mainstream dataset) will have difficulties performing evenly on a Twitter dataset. Second, we will improve the NER annotations, since, currently, we provide only a few class labels for Named Entity tagging and these annotations are given by an *ad-hoc* set of rules.

Finally, we plan to integrate time related features into our modeling. Indeed, people are likely to post messages about a specific event, such as an earthquake, in the same time span as the event. This could enable clustering online posts around a shared event and help the disambiguation of some spatial entities by leveraging knowledge extracted from other posts in the same cluster.

ACKNOWLEDGMENTS

The work presented in this article was carried out within the framework of the RéSoCIO project co-financed by the French National Research Agency (ANR) under the grant ANR-20-CE39-001.

REFERENCES

Akter, S. and Wamba, S. F. (2019). "Big data and disaster management: a systematic review and agenda for future research". In: *Annals of Operations Research* 283.1, pp. 939–959.

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

Ardanuy, M. C. and Sporleder, C. (2017). "Toponym disambiguation in historical documents using semantic and geographic features". In: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pp. 175–180.

Arthur, R., Boulton, C. A., Shotton, H., and Williams, H. T. (2018). "Social sensing of floods in the UK". In: *PloS one* 13.1, e0189327.

Attardi, G. (2015). *WikiExtractor*. https://github.com/attardi/wikiextractor.

Auclair, S., Boulahya, F., Birregah, B., Quique, R., Ouaret, R., and Soulier, E. (2019). "SURICATE-Nat: Innovative citizen centered platform for Twitter based natural disaster monitoring". In: *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*. IEEE, pp. 1–8.

Blank, D. and Henrich, A. (2015). "Geocoding place names from historic route descriptions". In: *Proceedings of the 9th Workshop on Geographic Information Retrieval*, pp. 1–2.

Botha, J. A., Shan, Z., and Gillick, D. (2020). "Entity Linking in 100 Languages". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Association for Computational Linguistics, pp. 7833–7845.

Brando, C., Frontini, F., and Ganascia, J.-G. (2015). "Disambiguation of named entities in cultural heritage texts using linked data sets". In: *East European Conference on Advances in Databases and Information Systems*. Springer, pp. 505–514.

Bunescu, R. and Paşca, M. (Apr. 2006). "Using Encyclopedic Knowledge for Named entity Disambiguation". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, pp. 9–16.

Cardoso, A. B., Martins, B., and Estima, J. (2022). "A Novel Deep Learning Approach Using Contextual Embeddings for Toponym Resolution". In: *ISPRS International Journal of Geo-Information* 11.1, p. 28.

Cheng, Z., Caverlee, J., and Lee, K. (2010). "You are where you tweet: a content-based approach to geo-locating twitter users". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). "Improving efficiency and accuracy in multilingual entity extraction". In: *Proceedings of the 9th International Conference on Semantic Systems*, pp. 121–124.

Derczynski, L., Nichols, E., Erp, M. van, and Limsopatham, N. (2017). "Results of the WNUT2017 shared task on novel and emerging entity recognition". In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 140–147.

Derungs, C. and Purves, R. S. (2014). "From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus". In: *International Journal of Geographical Information Science* 28.6, pp. 1272–1293.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT (1)*. Association for Computational Linguistics, pp. 4171–4186.

Dohrn, H. and Riehle, D. (2011). "Design and implementation of the sweble wikitext parser: unlocking the structured data of wikipedia". In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pp. 72–81.

Fayjaloun, R., Gehl, P., Auclair, S., Boulahya, F., Guérin-Marthe, S., and Roullé, A. (2021). "Integrating strong-motion recordings and Twitter data for a rapid shakemap of macroseismic intensity". In: *International Journal of Disaster Risk Reduction* 52, p. 101927.

Geiß, J., Spitz, A., Strötgen, J., and Gertz, M. (2015). "The Wikipedia location network: overcoming borders and oceans". In: *Proceedings of the 9th workshop on geographic information retrieval*, pp. 1–3.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). "Learning to forget: Continual prediction with LSTM". In: *Neural computation* 12.10, pp. 2451–2471.

Ghaddar, A. and Langlais, P. (2017). "Winer: A wikipedia annotated corpus for named entity recognition". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 413–422.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Grace, R., Kropczynski, J., Pezanowski, S., Halse, S. E., Umar, P., and Tapia, A. H. (2017). "Social Triangulation: A new method to identify local citizens using social media and their local information curation behaviors." In: *ISCRAM*.

Gritta, M., Pilehvar, M., and Collier, N. (2018). "Which melbourne? augmenting geocoding with maps". In: *56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia, pp. 1285–1296.

Habib, M. B. and Keulen, M. van (2012). "Improving Toponym Disambiguation by Iteratively Enhancing Certainty of Extraction." In: *KDIR*, pp. 399–410.

Hunt, A. and Specht, D. (2019). "Crowdsourced mapping in crisis zones: collaboration, organisation and impact". In: *Journal of International Humanitarian Action* 4.1, pp. 1–11.

Kaufhold, M.-A. (2021). "Retrospective Review and Future Directions for Crisis Informatics". In: *Information Refinement Technologies for Crisis Informatics*. Springer, pp. 47–73.

Kim, J., Vasardani, M., and Winter, S. (2017). "Similarity matching for integrating spatial information extracted from place descriptions". In: *International Journal of Geographical Information Science* 31.1, pp. 56–80.

Li, J., Sun, A., Han, J., and Li, C. (2020). "A Survey on Deep Learning for Named Entity Recognition". In: *IEEE Transactions on Knowledge and Data Engineering* 34.1, pp. 50–70.

Logan, R., Liu, N. F., Peters, M. E., Gardner, M., and Singh, S. (July 2019). "Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5962–5971.

Martins, B., Anastácio, I., and Calado, P. (2010). "A machine learning approach for resolving place references in text". In: *Geospatial thinking*. Springer, pp. 221–236.

Melo, F. and Martins, B. (2017). "Automated geocoding of textual documents: A survey of current approaches". In: *Transactions in GIS* 21.1, pp. 3–38.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). "Pointer sentinel mixture models". In: *arXiv preprint arXiv:1609.07843*.

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

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). "Learning multilingual named entity recognition from Wikipedia". In: *Artificial Intelligence* 194, pp. 151–175.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to expect when the unexpected happens: Social media communications across crises". In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp. 994–1009.

Overell, S. and Rüger, S. (2008). "Using co-occurrence models for placename disambiguation". In: *International Journal of Geographical Information Science* 22.3, pp. 265–287.

Paris, P.-H., Abadie, N., and Brando, C. (2017). "Linking spatial named entities to the Web of data for geographical analysis of historical texts". In: *Journal of Map & Geography Libraries* 13.1, pp. 82–110.

Pennington, J., Socher, R., and Manning, C. D. (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). "Deep Contextualized Word Representations". In: *NAACL-HLT*. Association for Computational Linguistics, pp. 2227–2237.

Qadir, J., Ali, A., Rasool, R. ur, Zwitter, A., Sathiaseelan, A., and Crowcroft, J. (2016). "Crisis analytics: big data-driven crisis response". In: *Journal of International Humanitarian Action* 1.1, pp. 1–21.

Rajapaksha, P., Farahbakhsh, R., and Crespi, N. (2021). "BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits". In: *IEEE Access* 9, pp. 154704–154716.

Rasmussen, J. and Ihlen, Ø. (2017). "Risk, crisis, and social media: A systematic review of seven years? research". In: *Nordicom Review* 38.2, pp. 1–17.

Recchia, G. and Louwerse, M. (2013). "A Comparison of String Similarity Measures for Toponym Matching". In: *COMP 2013 - ACM SIGSPATIAL International Workshop on Computational Models of Place*. Pp. 54–61.

Santos, J., Anastácio, I., and Martins, B. (2015). "Using machine learning methods for disambiguating place references in textual documents". In: *GeoJournal* 80.3, pp. 375–392.

Speriosu, M. and Baldridge, J. (2013). "Text-driven toponym resolution using indirect supervision". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1466–1476.

Spitz, A., Geiß, J., and Gertz, M. (2016). "So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks". In: *Proceedings of the third international ACM SIGMOD workshop on managing and mining enriched geo-spatial data*, pp. 1–6.

Stanek, S. and Drosio, S. (2012). "A hybrid decision support system for disaster/crisis management". In: *Fusing Decision Support Systems into the Fabric of the Context*. IOS Press, pp. 279–290.

Starbird, K., Muzny, G., and Palen, L. (2012). "Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions." In: *ISCRAM*. Citeseer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is All you Need". In: *NIPS*, pp. 5998–6008.

Wang, J., Hu, Y., and Joseph, K. (2020). "NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages". In: *Transactions in GIS* 24.3, pp. 719–735.

Wang, Z. and Ye, X. (2018). "Social media analytics for natural disaster management". In: *International Journal of Geographical Information Science* 32.1, pp. 49–72.

Xu, C., Li, J., Luo, X., Pei, J., Li, C., and Ji, D. (2019). "DLocRL: A deep learning pipeline for fine-grained location recognition and linking in tweets". In: *The World Wide Web Conference*, pp. 3391–3397.

Yan, Z., Yang, C., Hu, L., Zhao, J., Jiang, L., and Gong, J. (2021). "The Integration of Linguistic and Geospatial Features Using Global Context Embedding for Automated Text Geocoding". In: *ISPRS International Journal of Geo-Information* 10.9, p. 572.

Zhang, J., Ahlbrand, B., Malik, A., Chae, J., Min, Z., Ko, S., and Ebert, D. S. (2016). “A visual analytics framework for microblog data analysis at multiple scales of aggregation”. In: *Computer Graphics Forum*. Vol. 35. 3. Wiley Online Library, pp. 441–450.