Spatial Entity Extraction Methods from Arabic Texts in the Context of Epidemiological Surveillance: A Comparative Study

Fatima Ezzahra El Houbri[1,2,3], **Najlae Idrissi**[1], **Mathieu Roche**[2,3], **Sarah Valentin**[2,3]
[1]IDC Team, FST, Sultan Moulay Slimane University, Beni-Mellal, Morocco
[2]CIRAD, UMR TETIS, F-34398 Montpellier, France
[3]TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
elhoubri.fatimaezzahra@usms.ac.ma, n.idrissi@usms.ma
{mathieu.roche, sarah.valentin}@cirad.fr

Abstract

The location of events plays a key role in epidemiological surveillance, as it allows epidemic outbreaks to be identified and their spread to be tracked. Several event-based surveillance systems rely on English translation before extracting spatial entities using pretrained models adapted to English. However, the step of translating Arabic texts prior to extraction can influence the quality of spatial entity recognition. This study compares two strategies for extracting spatial entities from Arabic texts related to epidemiological surveillance evaluated in terms of precision, recall, and F-measure. The first approach relies on machine translation using DeepL, Microsoft Azure, and Reverso, followed by the extraction of entities using the spaCy (small, medium, and large) and GLiNER models. The second strategy consists of directly extracting entities from Arabic texts by fine-tuning the AraBERT, AraELECTRA, ARBERT, and MARBERT models. The results indicate that the second approach offers superior performance, particularly with the AraBERT and AraELECTRA models, demonstrating the effectiveness of direct extraction on Arabic data compared to the translation-based strategy.

1 Introduction

Epidemiological information systems monitor various channels to detect and report emerging and re-emerging infectious diseases. These systems are classified into two main categories of surveillance: indicator-based surveillance (IBS), which uses official sources such as medical records, and event-based surveillance (EBS), which relies on informal sources such as social media or news articles on the internet. Despite its complex implementation, EBS stands out for its ability to provide early and timely alerts about health threats. This method is particularly suitable for remote areas or areas with inadequate health coverage, where unofficial information channels are the main source of information (Menya et al., 2024; Delon et al., 2025).

Recently, a lot of studies (Joshi et al., 2019) have focused on the use of textual data from unofficial sources for epidemic surveillance, particularly in the context of animal health surveillance. While organizations such as the World Organization for Animal Health publish official notifications, informal sources such as online news sites offer data that may be more timely. In practice, event-based epidemiological surveillance systems often rely on the automatic translation of text sources collected from different languages into a pivot language, usually English, to enable the extraction of epidemiological information using artificial intelligence and machine learning methods (Valentin et al., 2021).

This information includes spatial data. Accurate location determination is essential for assessing epidemic risks, enabling epidemiologists to identify potential outbreaks or monitor the spread of disease. This spatialization is a prerequisite for implementing appropriate control and protection measures and therefore requires high-quality translation. However, to our knowledge, the effect of translation prior to spatial entity extraction has not been evaluated for epidemiological surveillance, particularly for certain languages due to their stylistic and lexical diversity. These issues are particularly important for the Arabic language, which encompasses a wide variety of dialects, linguistic registers, and cultural nuances (Akan et al., 2019; Guellil et al., 2021). This complexity highlights the need to evaluate two methods/approaches of spatial entity extraction for detecting epidemiological events from Arabic-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 246–256
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_025

language data.

In this context, our work consists of comparing two Arabic data extraction strategies. The first strategy is based on a traditional translation-based approach, where Arabic data is translated into English before applying named entity recognition models trained in English (**Trans-NER-EN**). The second strategy consists of performing named entity recognition directly on Arabic data using models trained specifically for Arabic (**NER-AR**). Our objective is to evaluate the effectiveness of both strategies to identify the best approach to Arabic data extraction for accurately detecting relevant spatial entities for epidemiological surveillance.

The remainder of the article is organized as follows. Section 2 presents related work on text-based epidemiological surveillance systems and pre-trained language models for Arabic. Section 3 describes the corpus used and the process for annotating spatial entities. Section 4 details the first strategy evaluated, namely the machine translation approach followed by named entity extraction in English (Trans NER-EN). Section 5 presents the second strategy, which consists of directly extracting spatial entities from Arabic texts using pre-trained models for Arabic (NER-AR). Section 6 outlines the evaluation metrics used and the results obtained for both strategies. Section 7 discusses the results obtained. Finally, Section 8 concludes the article and outlines prospects for future work.

2 Related Work

Text-based health monitoring systems, particularly HealthMap (Freifeld et al., 2008) and PADI-web (Valentin et al., 2021), aim to detect and extract epidemiological information from textual sources, such as online news. These systems automate several stages of text data processing, from document collection (e.g., online news) to the extraction of epidemiological information that can be used for analysis through the incorporation of artificial intelligence and machine learning methods.

Historically, these systems have been designed primarily for English-language texts (Arsevska et al., 2018), which limits their ability to identify events reported in local languages, particularly Arabic. This limitation has led to the development of multilingual systems that increasingly target different languages to guaranty direct access to local information sources. These systems generally include a machine translation step into a target language, often English, using open-source or proprietary translation systems (e.g., Google Translate, Microsoft Translator), to enable the use of pre-trained models for downstream tasks such as classification, named entity recognition, and event localization. Although spatial entity recognition is essential for correctly identifying epidemics and tracking their spread, to our knowledge, the impact of translating Arabic texts before automatic extraction has not been evaluated for epidemiological surveillance.

Recently, pre-trained language models (PLM) for Arabic have demonstrated superior representation capabilities for capturing contextual information within sentences. In general, PLMs based on Transformer-type architectures allow for direct analysis of Arabic texts while taking into account their lexical, morphological, and syntactic specificities, which significantly improve various NLP tasks, including Arabic Named Entity Recognition, which involves detecting and classifying named entities in a given text, including spatial entities (El Moussaoui and Loqman, 2024; Qu et al., 2023).

AraBERT (Antoun et al., 2020a) is a pre-trained language model, based on the architecture of BERT (Devlin et al., 2019) in English and specially designed for the Arabic language. The model is based on bidirectional learning using masked language modeling (MLM) and next sentence prediction (NSP) tasks. In order to take into account the linguistic structure of Arabic, AraBERT uses external segmentation to break words down into roots, prefixes, and suffixes. AraBERT's performance was evaluated on three Arabic language comprehension tasks, specifically named entity recognition, question answering, and sentiment analysis.

AraELECTRA (Antoun et al., 2020b) is another pre-trained model for Arabic, based on the architecture of ELECTRA (Clark et al., 2020) in English. It contains a generator that predicts the original tokens from corrupted sentences and a discriminator that determines which tokens are masked and which ones are original. During fine-tuning, the input sentence is fed to the discriminator with an additional linear layer for downstream tasks. AraELECTRA was

evaluated on the ANERCorp dataset, and the results clearly show its superiority (Qu et al., 2023).

MARBERT (Abdul-Mageed et al., 2021), meanwhile, is a large-scale pre-trained masked language model for Arabic dialects and Modern Standard Arabic. Based on the BERT-base architecture, it uses a 128 GB corpus of Arabic tweets and removes the next-word prediction task. Pre-trained on a wide range of dialects, MARBERT achieves state-of-the-art performance on named entity recognition tasks, particularly on the TW-NER dataset (Darwish, 2013).

Finally, **ARBERT** (Abdul-Mageed et al., 2021) is a large-scale masked language model specifically focused on Modern Standard Arabic (MSA). ARBERT follows the same architecture as BERT base: 12 attention layers, each with 12 attention heads and 768 hidden dimensions, with a vocabulary of 100,000 WordPieces, representing approximately 163 million parameters. It was trained on a 61 GB corpus of Arabic text, equivalent to 6.2 B tokens.

Although several studies have used fine-tuning of these models to extract named entities, to our knowledge, none have focused on the extraction of spatial entities within epidemiological surveillance. In this context, we conducted a comparative study on the performance of these fine-tuned models, evaluated on Arabic data related to animal epidemiological surveillance. This approach was then compared to the strategy based on translation before extraction to determine the best method for retrieving these spatial entities.

3 Corpus and annotation

To evaluate and compare the performance of both strategies to extract spatial entities in English and Arabic for spatial entity recognition, we constructed two datasets from 96 Arabic news articles harvested by the PADI-web platform between February 2025 and February 2026. These datasets contain 328 manually annotated spatial entities, corresponding to 142 distinct spatial entity values. Spatial entities were manually annotated and extracted by the first author, a native Arabic speaker. Among these entities, there are 275 Absolute Spatial Entities (ASE). These entities can be found on a map or in a geographic database because they are direct references to specific, locatable geographic places, such as continents, countries, and cities. The remaining 53 entities are Relative Spatial Entities (RSE), which are specified in relation to other spatial entities utilizing topological relationships. Examples include " West of Mosul " and " northwestern Hungary ".

The first, PADI-Location-AR-EN[1], is a parallel Arabic-English dataset dedicated to spatial entities from epidemiological texts. It combines manually validated translations using geographic databases and machine translations performed by automatic translators, as well as normalization through the GeoNames database using a unique identifier. This corpus is developed to evaluate the impact of machine translation of spatial entities before extraction and the performance of English named entity recognition models for spatial entity recognition applied to translated texts. For more details on the corpus construction, annotation, extraction, and normalization process see our previous study (Houbri et al., 2026).

The second, PADI-Location-AR-IOB, is a dataset containing the same spatial entities as PADI-Location-AR-EN but annotated according to the IOB schema, in which the labels B-LOC and I-LOC mark the beginning and interior of spatial entities, respectively, while the label O refers to other tokens that do not correspond to any entity. This corpus is constructed to evaluate the performance of named entity recognition models in Arabic for spatial entity recognition.

4 Trans-NER-EN Strategy

This section aims to evaluate the preservation of spatial information across two classical and sequential steps in multilingual text processing: (1) the translation of Arabic locations into English and (2) the automated extraction of locations from the translated text.

4.1 Translation

A spatial entity extracted from texts in Arabic can undergo significant alterations when translated into another language, such as English, which can lead to a partial loss of meaning or distortion of

[1] https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/ZA1YUO

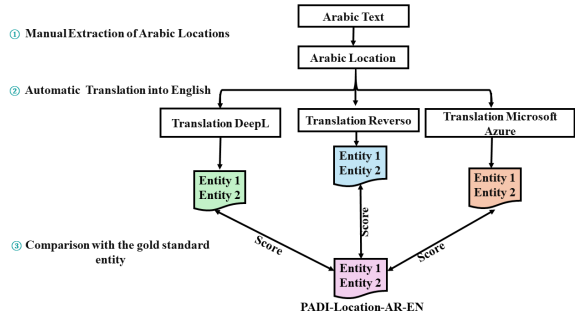

Figure 1: The Proposed Process for Translation Evaluation.

the spatial information (Almohish, 2024). In this context, computer-aided translation techniques play an essential role in identifying the most suitable machine translation system for converting spatial entities from Arabic to English. To this end, we are evaluating the performance of DeepL, Microsoft Azure, and Reverso by comparing their translations of spatial entities to the ground truth.

The three systems selected are briefly presented below, along with the reasons for their selection in our evaluation:

- **DeepL**[2]: One of the most renowned machine translation systems, based on convolutional neural networks (Laki and Yang, 2024). We chose DeepL because of its international reputation for excellent translations.
- **Microsoft Azure**[3]: The Microsoft Azure Translation API is a cloud-based machine translation service that you can use to translate texts and documents with a simple call to the REST API, which uses modern neural machine translation technology. This system is an operational service often integrated into text-mining tools, which makes its evaluation particularly relevant for epidemiological surveillance.
- **Reverso**[4]: Reverso is a suite of online applications and services using artificial intelligence to help translate, write, and improve language proficiency. It offers contextual translation, bilingual dictionaries, synonyms, conjugation, interactive grammar tools, as well as spelling and grammar correction based on neural machine translation [5]. We chose Reverso because it not only provides translations of individual words but also gives examples of paragraphs containing these translated words, enabling the translation to be checked in context.

Figure 1 shows the evaluation process for these three translation systems. The Arabic spatial entities, extracted and annotated manually from the 96 articles, served as a reference. We then automatically translated the sentences containing these entities using the three translators. The translated spatial entities were manually extracted from these texts, ensuring that they were translated in context and not in isolation. Finally, a comparison was made between the translated entities and those in our reference dataset. This comparison was done using two evaluation metrics: Exact Matching and Fuzzy Matching

4.2 Location extraction

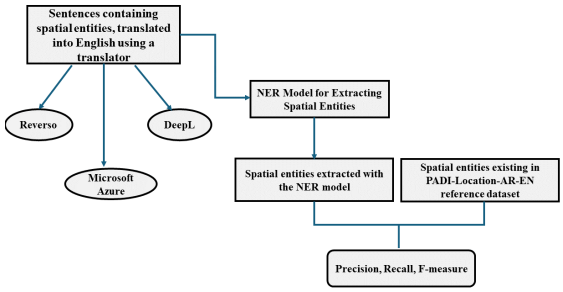

Figure 2: Spatial entities extraction pipeline based on extraction models, along with their evaluation.

As illustrated in Figure 2, to assess the quality of the named entity extraction models, we selected two named entity recognition models: spaCy[6] NER model, including its small[7], medium[8], and large[9] versions, and the GLiNER model (Zaratiana et al.,

[2] https://www.deepl.com/fr/translator
[3] https://learn.microsoft.com/en-us/azure/ai-services/translator/
[4] https://www.reverso.net/traduction-texte
[5] https://fr.wikipedia.org/wiki/Reverso_(logiciel)
[6] https://spacy.io/
[7] https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.8.0
[8] https://github.com/explosion/spacy-models/releases/tag/en_core_web_md-3.8.0
[9] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.8.0

2024), capable of identifying any type of entity using a BERT-type bidirectional transformer encoder. We used the models to extract spatial entities from the English-translated sentences, resulting in a set of extracted entities. Finally, we compared these entities to those in our reference dataset using exact matching, based on the classic metrics of precision, recall, and F-measure. It should be noted that the objective here is not to exhaustively evaluate all existing NER tools, such as NuNER (Bogdanov et al., 2024), UniversalNER (Zhou et al., 2024), or fine-tuned transformer-based models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), but rather to assess a small selection of representative and state-of-the-art models — some of which are often integrated into epidemiological surveillance platforms or text-mining tools — to demonstrate the usefulness of the PADI-Location-AR-EN dataset, which can serve as a resource for evaluating other NER tools in future work.

5 NER-AR Strategy

To evaluate the strategy of extracting spatial entities directly from Arabic data for epidemiological surveillance, as shown in Figure 3, we consider this approach as a named entity recognition (NER) task in Arabic. To do this, we fine-tuned several pre-trained language models for Arabic, including AraBERT, AraELECTRA, ARBERT, and MARBERT, to conduct a comparative study of their performance in spatial entity extraction.

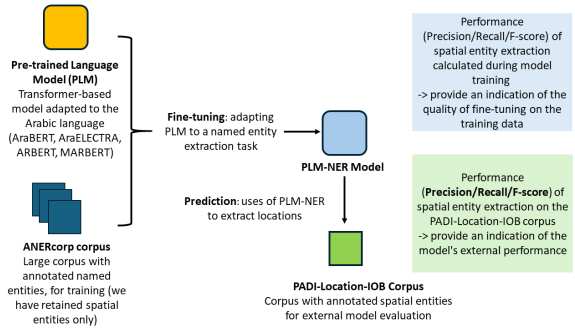

Figure 3: The proposed NER-AR Strategy on PADI-Location-AR-IOB

.

These models were trained on normalized training data containing only location entities from the ANERCorp corpus. The ANERCorp (Benajiba et al., 2007) corpus is a manually annotated Arabic corpus designed specifically for named entity recognition tasks. It consists of two subsets: train and test, and it contains over 150,000 tokens. Each token is associated with one of the following categories: person, place, organization, or other. It is crucial to note that in ANERCorp, location corresponds exclusively to absolute spatial entities, such as named toponyms (e.g., cities, countries, and regions). The corpus does not include or annotate spatial relative entities, which are not considered named entities within its annotation scheme.

During the normalization stage, we retained only annotations corresponding to spatial entities. More specifically, we retained the B-LOC and I-LOC labels, indicating the beginning and interior of spatial entities, respectively, as well as the O label for tokens that do not correspond to any entity. In total, the dataset contains **3,776 B-LOC**, **525 I-LOC**, and **120,801 O** labels. The Arabic PLMs were fine-tuned exclusively on the ANERCorp training split, using the following pre-trained checkpoints from the Hugging Face Hub: aubmindlab/bert-base-arabertv02 (AraBERT), aubmindlab/araelectra-base-discriminator (AraELECTRA), UBC-NLP/ARBERT (ARBERT), and UBC-NLP/MARBERT (MARBERT). Fine-tuning was performed using the Hugging Face Trainer with the AdamW optimizer, a learning rate of 5 × 10, a batch size of 16, gradient accumulation over two steps, a maximum sequence length of 256 tokens, weight decay of 0.01, six training epochs, a cosine learning rate scheduler with a 10% warm-up ratio, mixed-precision (FP16) training, and a fixed random seed of 42. Due to computational constraints, training was conducted with a single random seed; results may therefore be subject to some variance that was not quantified in this study. Precision, recall, and F-measure were computed using weighted averaging; since only one entity type (LOC) is annotated, this is numerically equivalent to a standard micro-average over entities.

We then evaluated the fine-tuned models on the PADI-Location-AR-IOB corpus, which is annotated with the same labels (B-LOC, I-LOC, and O) and contains exclusively spatial entities extracted from Arabic texts related to epidemiological surveillance. This dataset includes **328 B-LOC**, **115 I-LOC**, and **19,110 O** labels. No samples from the PADI-Location-AR-IOB corpus were used to update model weights or

to select the best checkpoint: no early stopping or best-model selection based on validation performance was applied, and training ran for a fixed number of epochs. Since our objective was to evaluate cross-domain generalization, no separate validation split was held out; test-set metrics were nonetheless logged periodically during training for monitoring purposes only, without influencing the training process or hyperparameter choices, and the fine-tuned models were evaluated directly on this independent test dataset at the end of training. Finally, each fine-tuned model was evaluated to conduct a comparative study of the models' performance. The results obtained were then compared with those of the Trans-NER-EN strategy to determine the best strategy for extracting spatial entities for the detection of epidemiological events from Arabic data. It should be noted that this comparison reflects two practical workflows rather than two strictly equivalent experimental settings, since the Arabic strategy relies on fine-tuned Arabic PLMs, whereas the Trans-NER-EN strategy uses off-the-shelf English NER systems after machine translation.

6 Evaluation

In this section, we introduce the evaluation metrics for translation and for the extraction of entities (from English-translated text and from Arabic text).

6.1 Evaluation Metrics for Translation

- **Exact Matching**: For each spatial entity translated by each of the translation systems, if the translation corresponds exactly to the character string present in our reference dataset, it is considered correct.
  The exact matching score corresponds to the percentage of spatial entities correctly translated by the translation system, calculated using Equation (1):

$$\text{Exact Matching Score} = \frac{\text{Number of correctly translated spatial entities}}{\text{Total number of spatial entities}} \quad (1)$$

- **Fuzzy Matching**: For each spatial entity translated by the translation systems, the translation is considered correct if it meets one of the following criteria. Fuzzy matching was performed manually according to predefined annotation guidelines. Each translated spatial entity was compared with the corresponding reference translation by the first author and subsequently reviewed by the co-authors to ensure consistency. Ambiguous cases were discussed collectively until a consensus was reached. No automatic semantic similarity metric was used. (1) exact match with the reference entity (e.g. 'Morocco' vs 'Morocco') (2) corresponds to an alternative name recognized in GeoNames or Google Maps (e.g. 'Fes' vs 'Fez') (3) differs from the reference entity by a single character operation according to the Levenshtein distance (e.g. 'Joumine' vs 'Jomine', by deleting the 'u' character), (4) has the same contextual meaning despite differences in wording while referring to the same geographic location (e.g. 'Northwestern China' vs. 'Northwest China'), or (5) is a reformulation or lexical rearrangement that retains the same meaning without modifying the referenced spatial entity (e.g. 'Oran Province' vs. 'Province of Oran').
  The fuzzy matching score corresponds to the percentage of spatial entities that satisfy at least one of these criteria, including exact matches, calculated according to Equation (2):

$$\text{Fuzzy Matching Score} = \frac{\text{Number of spatial entities meeting at least one of the criteria}}{\text{Total number of spatial entities}} \quad (2)$$

6.2 Evaluation Metrics for Extraction

To evaluate the performance of named entity extraction models, we used three metrics: precision, recall and F-measure. These metrics are calculated from true positives (TP), false positives (FP) and false negatives (FN), defined as follows:

- **True positives (TP)**: number of spatial entities correctly extracted by the model that appear in our reference dataset.
- **False positives (FP)**: number of spatial entities extracted by the model but not appearing in our reference dataset.
- **False negatives (FN)**: number of spatial entities not extracted by the model but appearing in our reference dataset.

Precision corresponds to the percentage of correct spatial entities extracted by the model, calculated

using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Number of spatial entities correctly extracted by the model}}{\text{Total number of spatial entities extracted by the model}} \quad (3)$$

Recall represents the percentage of spatial entities existing in our reference set that are found by the model and is calculated using formula (4):

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Number of spatial entities correctly extracted by the model}}{\text{Total number of spatial entities present in the reference dataset}} \quad (4)$$

A measure that incorporates both precision and recall is the harmonic mean of these two measures, known as the F-measure (formula (5)):

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

6.3 Results

6.3.1 Results of the translation

Table 1 shows the overall scores obtained by the translation systems evaluated. DeepL and Microsoft achieve the highest scores for exact matching, successfully translating 226 and 222 out of 328 entities, respectively. On the other hand, Reverso lags behind, with a score of 59.75%, reflecting a significantly lower level of accuracy and translation quality than DeepL and Microsoft Azure.
In terms of fuzzy matching, Microsoft Azure came out on top with a score of 86.58%, demonstrating a strong ability to produce translations that closely align with the intended meaning, even when not literal. DeepL follows closely with a score of 85.06%, confirming its consistency in translation quality. In contrast, Reverso scored lower at 74.39%, underlining weaker performance in preserving accurate translations.

Translator	Exact Matching Score (%)	Fuzzy Matching Score (%)
Reverso	196 / 328 = 59.75	244 / 328 = 74.39
DeepL	226 / 328 = **68.90**	279 / 328= **85.06**
Microsoft Azure	222 / 328 = **67.68**	284 / 328 = **86.58**

Table 1: Performance of the three translators for the translation of spatial entities in terms of exact and fuzzy match scores.

6.3.2 Results of entity extraction (from English-translated text)

According to the results presented in Table 2, Table 3, and Table 4, the spaCy small model achieves the lowest performance of the medium and large spaCy models, particularly in terms of recall. For example, it achieves an F-measure of 51.46% when used with Microsoft Azure, and 48.79% with DeepL. These results reflect the model's limitations in terms of complexity and ability to recognize entities in a variety of contexts, making it a less suitable choice for tasks requiring a high degree of language understanding. As shown in Table 3 and Table 4,

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	64.70	33.53	44.16
Microsoft Azure	**72.37**	**39.93**	**51.46**
DeepL	**70.93**	**37.19**	**48.79**

Table 2: Evaluation metrics of spatial entity extraction with spaCy's small model, according to the translations of the three translators.

the spaCy medium model delivers slightly lower performance than the large model, but is still superior to the small model. For example, when used with Microsoft Azure, it achieves an F-measure of 63.05%, while with DeepL, it achieves an F-measure of 51.04%.

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	64.28	35.67	45.88
Microsoft Azure	**71.65**	**56.30**	**63.05**
DeepL	**68.91**	**40.54**	**51.04**

Table 3: Evaluation metrics of spatial entity extraction with spaCy's medium model, according to the translations of the three translators.

Table 4 shows that spaCy's large model outperforms the small and medium models in spatial feature extraction, irrespective of which of the three evaluated translators is used. Its F-measure scores are higher than those of the other models. For example, with Microsoft Azure, it achieves an F-measure of 58.03%, while with DeepL, it achieves an F-measure of 54.53%.

Table 5 demonstrates the competitiveness of

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	62.28	38.10	47.27
Microsoft Azure	**73.70**	**47.86**	**58.03**
DeepL	**69.66**	**44.81**	**54.53**

Table 4: Evaluation metrics of spatial entity extraction with spaCy's large model, according to the translations of the three translators.

the GLiNER model, especially when paired with DeepL and Microsoft Azure translations. It surpasses the spaCy large model's scores with an F-measure of 67.96% with Microsoft Azure and

65.49% with DeepL. These results show that GLiNER offers improved spatial entity recognition capability, particularly in multilingual or translated contexts, and is a powerful alternative to traditional spaCy models. However, even with the best-performing approach (GLiNER), approximately 36.29% of locations are not extracted (63,71% recall), which is an important consideration for epidemiological surveillance applications.

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	61.70	53.04	57.04
Microsoft Azure	**72.82**	**63.71**	**67.96**
DeepL	**68.79**	**62.50**	**65.49**

Table 5: Evaluation metrics of spatial entity extraction with the GLiNER model, according to the translations of the three translators.

6.3.3 Results of entity extraction (from Arabic text)

We evaluated the performance of the fine-tuned AraBERT, AraELECTRA, ARBERT, and MARBERT models for extracting annotated locations from the EpiDataset2 dataset (Table 6).

Arabic PLM	Precision (%)	Recall (%)	F-measure (%)
AraBERT	**79.73**	**81.35**	**80.53**
AraELECTRA	**79.86**	**80.67**	**80.26**
ARBERT	79.5	78.45	78.98
MARBERT	79.6	67.68	73.16

Table 6: Performance of Arabic pre-trained language models (PLMs) for spatial entity extraction in terms of precision, recall, and F-measure. The two best values for each metric are shown in bold.

The obtained results demonstrate that the AraELECTRA model offers the best accuracy (79.86%), while AraBERT presents the best balance between accuracy and recall, with an F-measure of 80.53%. ARBERT performs slightly worse but remains competitive, and MARBERT achieves decent accuracy (79.6%) but lower recall (67.68%), which reduces its F-score (73.16%). Overall, these performances confirm the effectiveness of fine-tuning pre-trained models on standard Arabic for direct extraction of spatial entities.

7 Discussion

The results obtained show that the performance of absolute and relative spatial entity detection depends heavily on the choice of translator and extraction model used. Indeed, the quality of the translation directly influences the ability of the NER model to correctly identify spatial entities. The aim is to determine which translator provides the most accurate results while faithfully preserving the original spatial meaning. It should be noted that the aim of this study is not to comprehensively evaluate all current translation tools, especially since many other AI-based systems, such as Google Translate and others, may offer competitive performance. Instead, we selected a representative panel of three translators to provide an initial direction and demonstrate the usefulness of the manually created set in supporting this type of evaluation. Future work can easily extend this approach to other translation systems. Furthermore, the inclusion of Microsoft Azure is particularly significant. This allows us to assess the reliability of the translation tool currently used for epidemiological surveillance. In addition, when analyzing errors, translators generally do not make mistakes with the same spatial entities.

In the Trans-NER-EN strategy based on translation, even when combining the best translation system with the most effective extraction model (GLiNER), around 36.29% of spatial entities are not extracted, representing a recall of 63.71%, which is important to take into account for epidemiological surveillance applications. Extrapolating, this means that 36.29% of locations, and therefore potentially outbreaks, are ”missed”. There is therefore considerable room for improvement. An important perspective would be to compare the performance obtained in this study with that reported for NER models when fine-tuned on English training data, to assess the extent to which translation from Arabic impacts their performance. Furthermore, the NER-AR strategy, which relies on the direct extraction of spatial entities from Arabic texts, appears to offer better performance for the identification of spatial entities, highlighting the relevance of an approach that is linguistically adapted to the Arabic context. Encoder-based PLMs such as AraBERT and AraELECTRA are specifically designed for sequence labeling tasks like NER and can be efficiently fine-tuned on small specialized corpora such as ours (328 spatial entities) with limited computational resources. In contrast, LLMs such as GPT-4 and LLaMA are generally used in few-shot or zero-shot settings without fine-tuning, which makes them less suitable for specialized and small-scale corpora

such as those encountered in epidemiological surveillance. However, a direct comparison with the translation-based strategy remains to be established, as the spatial entities extracted in Arabic have not yet been normalized or translated into English. In this context, relative spatial entities (RSEs) represent a major challenge in terms of normalization. This can be explained by the very nature of these entities: they do not correspond to officially recognized administrative units but rather to directional or contextual descriptions of known places, such as ”northwestern China” or ”west of Mosul.” The next step will therefore be to evaluate the geocoding of the spatial entities extracted using both strategies in order to assess the consistency between the two methods. It would also be interesting to examine the feasibility of direct geocoding (Mubarak and Hassan, 2021), which involves converting spatial entities into geographic coordinates, from Arabic entities, without going through translation, which would reduce dependence on machine translation tools and improve the overall accuracy of the process.

8 Conclusion

In this work, we conducted a comparative study comparing two spatial entity extraction strategies for detecting epidemiological events from Arabic data, using precision, recall, and F-measure as evaluation metrics. The first, known as the traditional strategy, involves translating Arabic texts into English before applying named entity recognition models trained in English (Trans-NER-EN) to extract spatial entities from the translated texts. The second strategy (NER-AR) consists of directly extracting spatial entities from Arabic data using pre-trained models adapted to this language. The objective of this study was to evaluate the effectiveness of these two approaches to identify the most suitable strategy for accurately extracting relevant spatial entities for epidemiological surveillance. The results show that the NER-AR strategy gives the best results in terms of precision, recall, and F-measure, particularly thanks to the application of fine-tuning on pre-trained models for Arabic. In contrast, the translation-based Trans-NER-EN strategy performs less well, highlighting the need for high-quality translation and high-performance pre-trained models for the target language to extract translated entities correctly. However, these results are based on a relatively small dataset drawn from a single surveillance source/domain, and should not be taken to imply that the same performance ranking will necessarily hold across other Arabic dialects, data sources, or types of epidemiological events. Future work could explore the extraction of other types of epidemiological entities, such as hosts and symptoms, and evaluate the performance of translation-based approaches using newer translation tools based on large language models (LLMs). It would also be valuable to extend the comparison to additional Arabic pre-trained language models, such as CAMeLBERT, as well as multilingual models, to provide a broader benchmark. Furthermore, future studies could complement the evaluation with translation quality metrics such as BLEU and ChrF++ (Kumar et al., 2026), perform more detailed error analyses according to spatial entity type (e.g., absolute versus relative spatial entities), and report uncertainty estimates or statistical significance analyses to better assess the robustness of the observed differences. Finally, qualitative examples comparing the outputs of the two extraction strategies would help illustrate their respective strengths, limitations, and practical relevance for epidemiological surveillance.

Acknowledgments

The authors thank Julien Rabatel for using the Microsoft Azure technology linked with PADI-web to translate phrases into English.
Under the ”PhD Associate Scholarship – PASS” initiative for doctorate student monitors in the digital field by 2027—National Strategy Morocco 2030—the CNRST is supporting this work.
The MOOD (Monitoring Outbreaks for Disease surveillance in a data science context) project, DGAL, and the French National Research Agency under the Investments for the Future Program: ANR-16-CONV-0004 (DigitAg) have provided funding for this work.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Md Faruquzzaman Akan, Md Rezaul Karim, and Abdullah Mohammad Kabir Chowdhury. 2019. An analysis of arabic-english translation: Problems and prospects. *Advances in Language and Literary Studies*.

N. R. I. Almohish. 2024. Spatial and diachronic challenges encountering literary arabic translators of the classical arabic literature into english. *Research Journal in Advanced Humanities*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Araelectra: Pre-training text discriminators for arabic language understanding. *ArXiv*, abs/2012.15516.

Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn De Goër de Hervé, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. 2018. Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system. *PLoS ONE*, 13(8). Publisher: Public Library of Science.

Yassine Benajiba, Paolo Rosso, and José-Miguel Benedí. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 148–159. Springer, Berlin, Heidelberg.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.

François Delon, Gabriel Bédubourg, Léo Bouscarrat, Jean-Baptiste Meynard, Aude Valois, Benjamin Queyriaux, Carlos Ramisch, and Marc Tanti. 2025. Infectious risk events and their novelty in event-based surveillance: new definitions and annotated corpus. *Language Resources and Evaluation*, 59(1):277–295.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Taoufiq El Moussaoui and Chakir Loqman. 2024. Advancements in arabic named entity recognition: A comprehensive review. *IEEE Access*.

Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. 2008. Healthmap : Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15:150–157.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, pages 497–507.

Fatima Ezzahra El Houbri, Najlae Idrissi, Mathieu Roche, and Sarah Valentin. 2026. Padi-location-aren: A normalized arabic–english spatial entity dataset for epidemiological surveillance. *Data in Brief*.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2019. Survey of text-based epidemic intelligence: A computational linguistic perspective. *ACM Computing Surveys*.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2026. Evaluating extremely low-resource machine translation: A comparative study of chrf++ and bleu metrics.

László János Laki and Zijian Győző Yang. 2024. Improving machine translation capabilities by fine-tuning large language models and prompt engineering with domain-specific data. *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*, 2014:1-5.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Edmond Menya, Roberto Interdonato, Dickson Owuor, and Mathieu Roche. 2024. Explainable epidemiological thematic features for event-based disease surveillance. *Expert Systems with Applications*, 250:123894.

Hamdy Mubarak and Sabit Hassan. 2021. UL2C: Mapping user locations to countries on Arabic Twitter. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 145–153, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*, 36:943–959.

Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. 2021. Padi-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, 13.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. *Proc. of the 2024 Conf. of the North American Chapter of the Ass. for Comp. Ling.: Human Language Technologies, NAACL 2024*, pages 5364–5376.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition.