Text Mining from Migration Narratives

David Ing[1] ✉, Fabien Delorme[1], Said Jabbour[1], Nelly Robin[2], and Lakhdar Sais[1]

[1] CRIL, CNRS - Université d'Artois, France {ing,delorme,jabbour,sais}@cril.fr
[2] Géographe, CEPED - IRD/Université Paris Cité, France nelly.robin@ird.fr

Abstract. The pervasive proliferation of textual information, combined with the swift advancement in data acquisition methods, has resulted in an overwhelming volume of data, making it challenging to uncover relevant patterns. Text mining is a crucial process for extracting noteworthy and non-trivial patterns, as well as valuable knowledge from extensive collections of textual data. In this paper, we present a step towards a text mining approach designed to harness migration narrative texts, those collected from interviews with migrants during their journeys in English and French. Our contributions can be summarized as follows: (1) We first collaborate with experts in Humanities and Social Sciences (HSS) to annotate the essential domain concepts, their related terms, and the locations mentioned in those narratives. (2) To automatically extract such related terms embedded in the narratives, we propose adapting a set expansion algorithm in a weakly supervised manner using a tiny set of annotated terms. We then evaluate the proposed algorithm by comparing its output terms to those annotated by experts. (3) We utilize some existing frameworks to automatically identify locations crossed by migrants, followed by a disambiguation model to precisely pinpoint them on a map. To evaluate the proposed systems, we conduct the experiments by comparing their recognized locations and disambiguated locations to those annotated by experts. (4) We design a tool to visualize the itineraries of those locations on a map, enabling the observation of migration routes. Our discussions with HSS experts reveal that our proposed approach assists their analyses by automatically retrieving pertinent terms and drawing itineraries of migrants on a map, enabling a comprehensive understanding of their construction.

Keywords: Text Mining · Humanities and Social Sciences · Migration Narratives · Migration Routes.

1 Introduction

More than 80 percent of the data generated and collected by organizations is in the form of unstructured data [10]. This type of data manifests in various forms across different fields, including email bodies, press releases, contracts, medical records, speech-to-text snippets, and more. Given its ubiquitous nature, unstructured data provides a rich source of information. For example, the migration narrative texts in our study are collected from interviews with migrants

during their journeys. This type of data offers valuable insights that enable experts in Humanities and Social Sciences (HSS) to analyze migration routes and gain a profound understanding of migration phenomena. However, extracting, acquiring, and formalizing knowledge from unstructured text data effectively and promptly presents several challenges. Firstly, the sheer volume of unstructured text makes manually extracting critical concepts or terms arduous. Secondly, a deep understanding of the specific domain is crucial for extracting the desired information. Lastly, unstructured text often contains noise, such as misspellings, run-on words, additional whitespace characters, and abbreviations. This paper is part of an innovative ANR HYCI (Hyper-lieux, Crises, Migrations et Inégalités) Project (ANR-22-CE55-0010) that aims to extract complex knowledge (migration routes formulated as attributed graphs) from migrants' narrative texts, focusing on their journeys towards Western Europe. The complexity of these graphs relates to the attributes that label both the *nodes* and *arcs*. More specifically, *nodes* provide information about the resources associated with migration, such as financial support (e.g. aid from family, work experiences, etc.), accommodation (e.g. legal or illegal housing, etc.) and environments (e.g. sea, forest, desert, etc.), whereas *arcs* are annotated with various types of information, including means of transport, cost, and duration. A migrant's journey transcends mere spatial and temporal dimensions; it is characterized by various events, resources, constraints, and opportunities. The unique nature of each journey, along with the richness and diversity of migratory experiences, poses significant challenges in terms of modeling and representing the concepts within this domain and the relationships that connect them.

To address the aforementioned challenges, this paper presents a step towards a text mining technique designed to harness the wealth of migration narrative texts articulated by migrants in two languages: English and French. This approach represents an initial step in unraveling the intricate structure of the migratory experiences highlighted above. Our contributions can be summarized as shown in Fig. 1 and presented as follows:

1. In Fig. 1(a), we first collaborate with experts in HSS to annotate domain concepts and their related terms based on a narrative corpus and a list of extracted terms. In Fig. 1(b), all the locations mentioned in such narratives are annotated and then disambiguated using Wikidata [31] by experts to generate the corpus of annotated and disambiguated locations.
2. In Fig. 1(c), we propose adapting a set expansion algorithm to extract the related terms embedded within the narratives in a weakly supervised manner using only a tiny set of annotated terms. We evaluate the proposed algorithm by comparing its output terms to the annotated terms.
3. In Fig. 1(d), we employ some existing models to identify those locations, and disambiguate them using a pretrained model as shown in Fig. 1(e). We conduct the evaluations by comparing the recognized locations and disambiguated locations of the proposed systems to those annotated by experts.
4. In Fig. 1(f), we design a tool to visualize the itineraries of those locations.

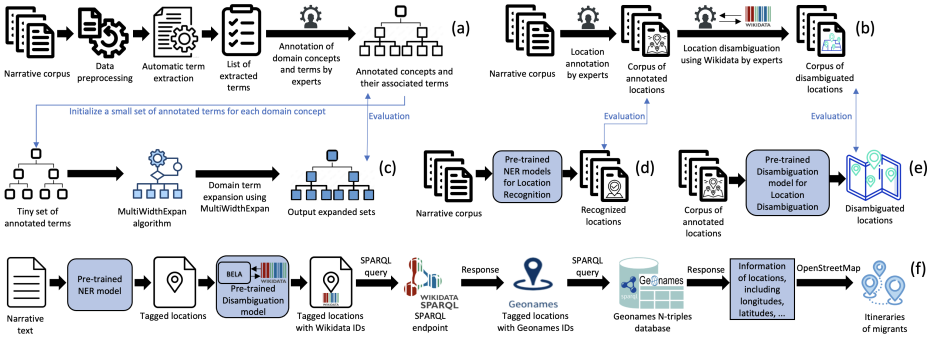

Fig. 1. The overall proposed framework for text mining from migration narratives.

The rest of this paper is organized as follows. Section 2 discusses related works on text mining in HSS. Section 3 presents the description of migration narratives. Section 4 describes our contributions for mining those narratives. Section 5 reports the evaluations of the proposed methods. Section 6 delineates the process of visualizing the itineraries of migrants on a map and discuss it with experts. Finally, Section 7 concludes the paper and addresses future works.

2 Related Work

This section briefly outlines the main approaches related to our works, focusing mainly on text mining in HSS and its use for migration narrative texts analysis.

In [17], the author discusses the state and the future of computational text analysis in sociology for social science by summarizing how different text mining tools have been used by sociologists in various analytical tasks across research questions and methodological traditions. The author describes five families of computational methods used in recent research: dictionary methods, semantic and network analysis tools, language models, unsupervised, and supervised machine learning. Among them, we tried the unsupervised clustering to identify the interesting concepts in our small migrant's narrative corpus. Unfortunately, the clusters that are supposed to delineate groups of terms enabling the emergence of concepts contain many noisy terms, making their identification challenging. Thus, we propose an alternative solution which we will describe in the sequel.

In [35], using a text mining approach/sentiment analysis on Twitter data, the authors investigate the public opinions and sentiments towards the Syrian refugee crisis. In [15], the authors analyze blogs to study shift in narratives within the blogosphere towards refugees or migrants during the migrant crisis in Europe. They use named-entity extraction to identify different topics and themes, followed by the use of targeted sentiment analysis to study the shift in narratives toward migrants in the blogosphere. In the two previously mentioned studies focusing on Twitter and blogosphere texts, the authors examined public sentiment regarding Europe's major refugee crises. However, our study breaks

new ground by analyzing narratives directly from migrants (in their own words during the interviews) to uncover the key concepts and their associated terms. Moreover, we extract and visualize the routes taken by the migrants on a map.

3 Migration Narratives Acquisition and Description

Studies on minor migrants often use narratives to convey representations of the migratory experience; narrative texts are collected from interviews conducted during their migratory journeys. These interviews do not occur in the familiar surroundings of the minors' villages or in their country of origin, nor in the host country where they may have found security, but rather in the transit points along the migration route. The minors are in a situation of waiting and in great uncertainty. This difficult and sometimes hostile environment requires a relationship of trust, which has been established through a period of time, the use of the language spoken by the minors, and the guarantee of anonymity for their testimonies. This raises the question of tensions between the researchers' goals in their field, professional ethics and moral responsibilities. The interviewers ensured that each life story is collected in accordance with social science ethical rules. The life stories are anonymized without causing any privacy concerns. Researchers from transit countries (Algeria, Mali, Morocco, Niger, Senegal) and local associations, trained in scientific data collection, are involved in the process. More specifically, life stories are collected from sub-Saharan minors in Niger (Agadez and Arlit), Algeria (Adrar, Tamanrasset and Maghnia), Senegal (Dakar, Mbour and Ziguinchor), and Morocco (Oujda and Rabat). They focus on three corridors along the *Trans-Saharan* routes [23]: (1) *Algerian-Malian corridor*: minors arrive from West African countries (Ivory Coast, Guinea-Bissau, Guinea, Mali, Nigeria and Senegal), and from Central Africa (the Democratic Republic of Congo (DRC)). (2) *Algerian-Nigerian corridor*: minors come from Nigeria and the DRC. (3) *Morocco-Senegal corridor*: the countries of origin of the migrants are from West Africa, including Senegal. In their narratives, minors tell their stories using a selection of places, means of transports, accommodations, words that designate objects and represent subjective experience.

Using similar methodology, other migrants' narratives are collected through the *Balkan* route [2]: one of the main migratory pathways into Europe. Transit corridors from Bulgaria, North Macedonia and Serbia, as well as through Albania and Montenegro, via Bosnia and Herzegovina, became one of the most travelled mixed migration routes in the Western Balkans. The migrants present along this route generally come from Syria, Pakistan, Afghanistan, Irak, Iran and Turkey.

4 Text Mining from Migration Narratives

This section presents our framework for mining migration narrative texts in three steps: (1) Narratives preprocessing, term extraction, and annotation, (2) Automatic domain term expansion, and (3) Location recognition and disambiguation.

4.1 Narratives Processing, Term Extraction & Concept Annotation

Narratives preprocessing is an initial step before extracting high-quality terms. It consists of basic natural language processing (NLP) tasks. For our narratives, we replace multiple consecutive punctuations with a single punctuation and multiple consecutive whitespace characters with a single whitespace character. We also remove all non-ASCII characters except accent characters, and some specific characters like š, đ, č, ć, ž for both lowercase and uppercase because those characters can be the name of certain places or locations (e.g. Šid, Krnjača, etc.).

To automatically extract terms, we first discussed with HSS experts and reached a consensus that single-word and two-word terms containing part-of-speech as nouns are relevant. Therefore, we use an approach that combines linguistic and statistical information to extract terms from both types of narratives.

Linguistic part. Inspired by [9], this approach consists of three core elements: (1) *Part-of-Speech (PoS)*: involves tagging the entire corpus with grammatical categories (e.g. noun, verb, adjective, etc.). (2) *Linguistic Filter*: filters desired terms from the resulting PoS tags, permitting only specific grammatical string for extraction. In our case, candidate terms are retained if they respect syntactic patterns such as: *Noun* and *Noun Noun*. *Noun* can be "*NN*", "*NPS*", "*NNP*", etc. (3) *stop-list*: a list of words excluded as terms in a domain.

Statistical part. We use Term Frequency-Inverse Document Frequency (TF-IDF) [25], a numerical statistic that scores terms in a text to indicate how important a term is to a document in a corpus. TF-IDF is calculated as follows:

$$TF\text{-}IDF(e, d, \mathcal{D}) = tf(e, d) * idf(e, \mathcal{D}),$$

$$tf(e, d) = \frac{f(e, d)}{\max\{f(w, d) : w \in d\}}, \quad idf(e, \mathcal{D}) = \log\left(\frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : e \in d\}|}\right),$$

where e is a term or entity (note that throughout the paper, we will use the words "term" and "entity" interchangeably), d is a document, $\mathcal{D}$ is a set of documents (i.e. a corpus), $f(e, d)$: number of occurrences of e in d, $tf(e, d)$: term frequency of e in d, and $idf(e, \mathcal{D})$: inverse document frequency of e in $\mathcal{D}$.

Following discussions with HSS migration experts on the previously extracted terms and narrative texts, we identified four essential and less sensitive domain concepts, along with a general concept, for both types of narratives as follows:

- **Family or Friends**: Family members or friends who financially support migrants in starting and continuing their journeys. They can also be companions during the journey (e.g. parents, siblings, friends, etc.).
- **Accommodation**: Locations or items where migrants can stay during a stopover (e.g. house, room, bed, etc.).
- **Means of transport**: Modes of transport or infrastructure utilized by migrants during their journeys (e.g. foot, plane, bus, road, etc.).
- **Environment**: Identifies the spatial and natural environments traversed by migrants during their journey (e.g. sea, forest, jungle, desert, etc.). These indicators highlight the difficulties, dangers, and risks of the migration routes.

Table 1. Statistics of migration narrative texts.

Narrative type : $\|\mathcal{D}\|$	*avg-tok-len*	$\|V\|$	$\#LOC$	Domain concept : $\#Entities$
English: 29	4148	1407	3039	**Accommodation:** 16, **Family or Friends:** 31, **Means of transport:** 28, **Environment:** 18
French: 108	1309	1719	2553	**Accommodation:** 20, **Family or Friends:** 35, **Means of transport:** 30, **Environment:** 14

- **Location**: A general concept denoted as a geographical location where migrants make a stopover, or a final destination city or country where they arrive (e.g. Greece, Istanbul, Gao, etc.). Extracting such locations is essential for HSS experts to clearly analyze the migration routes.

Table 1 shows the statistics of the narrative texts considered in this paper, including the total number of narratives ($|\mathcal{D}|$) in English and French, the average number of words (*avg-tok-len*), the total number of extracted terms ($|V|$), where V is a list of extracted terms, the total number of annotated and disambiguated locations ($\#LOC$), and the total number of associated terms for each domain concept ($\#Entities$). Note that $\#LOC$ and $\#Entities$ are annotated by experts.

4.2 Automatic Domain Term Expansion

In this section, we aim to automatically identify all the associated terms for each domain concept. Explicitly, our goal is to initialize a small set of annotated terms for each concept, and then find a complete set of terms belonging to the same concept. To achieve this, we use an algorithm called "set expansion". It refers to a technique of expanding a given partial set of seed entities into a more complete set of entities that belong to the same semantic class [33]. Previous works for solving this task include *Google Set* [30], *SEAL* [33], and *Lyretail* [6]. Other studies [13, 26, 28, 32] and Width Expansion in HiExpan [27] are proposed in a *corpus-based* setting where sets are expanded through an offline process.

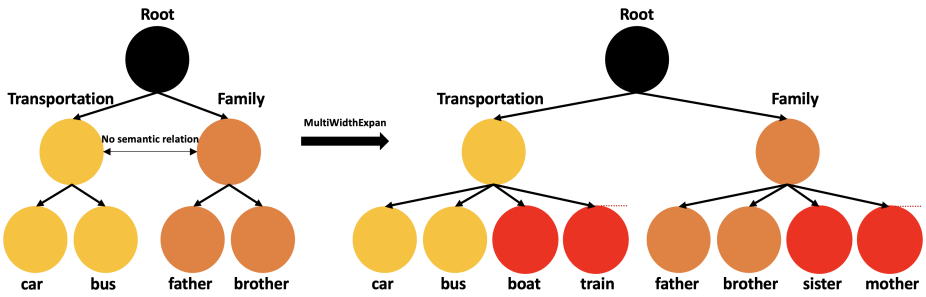

Fig. 2. Example of MultiWidthExpan with two different concepts.

Since our work does not exploit external resources and relies solely on small corpora of migration narratives (see Table 1) articulated naturally by migrants,

plus a tiny set of annotated terms (seed set), we propose adapting the Width Expansion in HiExpan [27]. It is a new and modified version of SetExpan [26], which additionally incorporates the term embedding feature and enhances the entity type features. The input for WidthExpan includes three parts: (1) a list of extracted terms V, (2) a set of documents $\mathcal{D}$, each document d contains tagged entities $e \in V$, and (3) a tiny set of annotated entities S. Given V and $\mathcal{D}$, we aim to expand S into a more complete set. For example (see Fig. 2), if a given seed set S of *Transportation* is {*car*, *bus*}, WidthExpan should return other related entities such as *boat*, *train*, etc. Note that we made some necessary changes to the original Width Expansion algorithm to handle both types of narratives, since the original one only allows to expand terms using English texts. Precisely, we execute the WidthExpan consecutively in each iteration, denoted as *MultiWidthExpan*. As shown in Fig. 2, each direct node from the root node does not belong to the same semantic class; they are simply the labeled concepts defined by experts to categorize related terms. Therefore, *MultiWidthExpan* executes the Width Expansion under each labeled concept successively (e.g. starting first with *Transportation*, then *Family*, and so on). We do not execute the Width Expansion separately for each concept, because at the end of each iteration, we globally optimize our *MultiWidthExpan* in case of conflict (the same term appears in multiple concepts), which we will describe in the sequel.

We now discuss the components of WidthExpan in detail, including the types of features, similarity measures, and the whole process of the Width Expansion.
Features. We use three types of features as follows:

1. ***skip-pattern***: Given a target term e_i in a sentence, one of its skip-pattern features is "w_{-1} ___ w_1" where w_{-1} and w_1 are two context words and e_i is replaced with a placeholder. For example, one skip-pattern of term "bus" in sentence "we went by bus to the border." is "by ___ to". We extract six skip-patterns of various lengths for one target term e_i in each sentence.
2. ***term embedding***: This feature captures the semantic similarity for each term. We have conducted preliminary experiments with various word embeddings and observed that the skip-gram [3] and continuous bag-of-words (CBOW) [12] models obtained the highest similarity scores for our English and French narratives, respectively. Thus, we utilize the pretrained skip-gram and CBOW models for the English and French corpora, respectively. Note that two-word terms (e.g. bus station) are concatenated using "_" before learning their embeddings. Since these models exploit subword information, they can handle out-of-vocabulary words, which is suitable for our case.
3. ***entity type***: Given {*car*, *bus*}, one of their common types is *vehicle*. For English, we obtain each entity's type by linking it to a knowledge base (KB) called Probase [34]. To our knowledge, no probabilistic taxonomy is available for French. Inspired by [20], we propose to query a French Masked Language Model (MLM) called CamemBERT[3] [18] with cloze-style prompts [29] to obtain the types of entities for creating a KB for French. Precisely, we design a specific prompt that combines an entity and a short sentence as follows:

[3] https://huggingface.co/almanach/camembert-base

"[ENTITY] est le type de [MASK].",

where [ENTITY] represents an entity $e \in V$, and [MASK] represents a hidden word to be predicted as the type of e by the MLM. For example, to find types of the term "camion" in French, we submit a specific prompt: "camion est le type de [MASK]". CamemBERT will return the following results:

```
[(véhicule, 0.34907031059265137),
(camion,    0.1406002938747406),
(voiture,   0.11208264529705048),
...]
```

Similarity Measures. We want to compute the sibling similarity between two entities e_1 and e_2 denoted as $sim_{sib}(e_1, e_2)$. First, the TF-IDF weight [24] between an entity e and a skip-pattern sk is calculated as follows:

$$f_{e,sk} \;=\; \log\big(1 + X_{e,sk}\big) \left[\log|V_{sk}| \;-\; \log\Big(\sum_{\substack{e' \in V_{sk} \\ e' \neq e}} X_{e',sk} \Big) \right], \tag{1}$$

Similarly, the association weight between an entity e and a type tp is calculated as follows:

$$f_{e,tp} \;=\; \log\big(1 + C_{e,tp}\big) \left[\log|V_{tp}| \;-\; \log\Big(\sum_{\substack{e' \in V_{tp} \\ e' \neq e}} C_{e',tp} \Big) \right], \tag{2}$$

Then, we can calculate the weighted Jaccard similarity [16] of two sibling entities using *skip-pattern* features as follows:

$$sim^{sk}_{sib}(e_1, e_2|SK) = \frac{\sum_{sk \in SK} min(f_{e_1,sk}, f_{e_2,sk})}{\sum_{sk \in SK} max(f_{e_1,sk}, f_{e_2,sk})}, \tag{3}$$

Similarly, the weighted similarity of two sibling entities using *type* features is calculated as follows:

$$sim^{tp}_{sib}(e_1, e_2|TP) = \frac{\sum_{tp \in TP} min(f_{e_1,tp}, f_{e_2,tp})}{\sum_{tp \in TP} max(f_{e_1,tp}, f_{e_2,tp})}, \tag{4}$$

where $X_{e,sk}$ is the raw co-occurrence count between entity e and skip-pattern sk. $V_{sk} \subseteq V$ is a list of entities contained in noun phrases (NPs), and $|V_{sk}|$ denotes its size. $C_{e,tp}$ is the confidence score of entity e having type tp. $V_{tp} \subseteq V$ is a list of entities that have types in the KB, and $|V_{tp}|$ denotes its size. SK and TP are selected sets of skip-pattern and type features associated with entities in V_{sk} and V_{tp}, respectively. Note that $V_{sk} \neq V_{tp}$, since some entities are not contained in NPs but have types, and vice versa. For example, a term "refugee" has no types in Probase. Therefore, it is not included in V_{tp}.

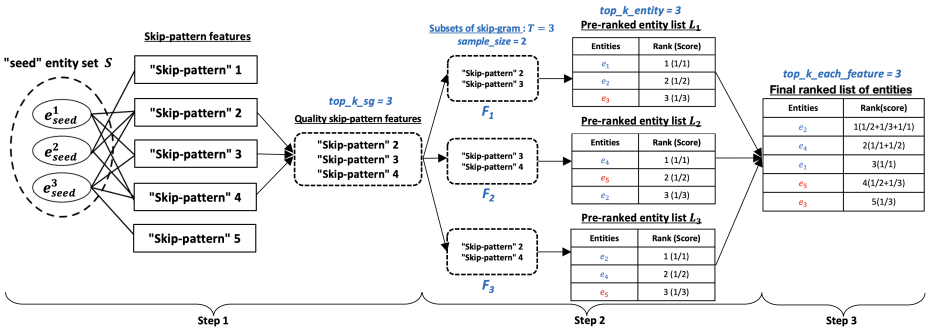

Fig. 3. Example of WidthExpan process using *skip-pattern* features.

Finally, the similarity between two entities based on their *embedding* features via cosine similarity is calculated as follows:

$$sim_{sib}^{emb}(e_1, e_2) = \frac{e_1.e_2}{\|e_1\|_2 * \|e_2\|_2}, \tag{5}$$

where $\| \ \|_2$ is the 2-norm, $e_1.e_2$ is a dot product of entity e_1 and entity e_2.

We combine the three similarities mentioned above by concluding that a good pair of sibling entities can appear in similar contexts, share similar embedding, and have similar types. Thus, the sibling similarity is calculated as follows:

$$sim_{\text{sib}}(e_1, e_2) \ = \ \sqrt{\left(1 + sim_{\text{sib}}^{sk}(e_1, e_2|SK)\right)\left(1 + sim_{\text{sib}}^{tp}(e_1, e_2|TP)\right) sim_{\text{sib}}^{emb}(e_1, e_2)} \tag{6}$$

The confidence score of entity e_1 belonging to S, $|S|$ is the total number of seed entities, is calculated as follows:

$$conf(e_1) = \frac{\sum_{e_2 \in S} sim_{sib}(e_1, e_2)}{|S|} \tag{7}$$

Reciprocal Rank. The Reciprocal Rank (RR) determines the reciprocal of the rank at which the initial pertinent document appears [7]. RR is 1 if a pertinent document is found at the first position, $\frac{1}{2}$ if it is found at the second position and so on. Here, a document represents an entity. Therefore, we can calculate the RR of an entity e according to its ranking position, denoted as $RR(e) = \frac{1}{r_e}$, where r_e is the ranking position of the entity e.

WidthExpan Overall Process. Given a seed set S and a list of extracted entities V, we follow [27], since the feature space is huge (i.e. many skip-pattern features are noisy) and V is noisy (i.e. many entities in V are irrelevant to S).

To obtain candidate entities using *skip-pattern* features (see Fig. 3), we apply three steps: (1) To score each skip-pattern feature, we accumulate its strength with entities in S by calculating $score(sk) = \sum_{e \in S} f_{e,sk}$, and select top_k_sg skip-pattern features with highest scores. (2) We use sampling without replacement method to generate T subsets of skip-pattern features F_t, where

$t = 1, \ldots, T$ and each F_t has a fixed size of *sample-size* features. For each F_t, we select candidate entities from V_{sk} if they have associations with skip-pattern features in F_t. We then calculate similarity scores between the selected candidate entities and the seed entities in S (see equation (3)), arrange them based on their scores in descending order, and compute their reciprocal ranks (RRs). (3) From each list L_t, we select *top_k_entity* candidate entities, aggregate their RRs, arrange them in descending order in the final ranked list, and select the *top_k_each_feature* candidate entities.

Similarly, to obtain candidate entities using *entity type* features, we apply three steps: (1) To score each type feature, we accumulate its strength with entities in S by calculating $score(tp) = \sum_{e \in S} f_{e,tp}$, and select *top_k_type* type features with highest scores. (2) We select candidate entities from V_{tp} if they have associations with these *top_k_type* type features. (3) We calculate similarity scores between the selected candidate entities and the seed entities in S (see equation (4)), arrange them based on their scores in descending order, compute their RRs, and select the *top_k_each_feature* candidate entities.

To obtain the candidate entities using *term embedding* features, we apply two steps: (1) We combine the previously selected candidate entities of *skip-pattern* and *entity type* features. (2) We calculate similarity scores between these combined candidate entities and the seed entities in S (see equation (5)), arrange them based on their scores in descending order, calculate their RRs, and select the *top_k_each_feature* candidate entities.

Finally, we aggregate the reciprocal ranks of the selected candidate entities using these three types of features, arrange them in descending order, and select the *top_k_expand* candidate entities to expand S. For each concept, we set a maximum of *top_k_per_concept* terms. The confidence score of each candidate entity belong to S is calculated using equation (7).

Conflict Resolution and Global Optimization. At the end of each iteration, since the supervision signal from each seed entity S is very weak (i.e. only a few annotated terms are provided), we need to ensure that candidate terms introduced in early iterations are high-quality and will not mislead the expansion process later on. When a term appears in multiple concepts during the expansion process, we encounter a "conflict" and aim to resolve it by finding the best concept for that term. Given a set of conflicting terms C, we apply two rules:

1. If a term $c \in C$ is in a seed entity of any concept, we retain c in that concept and remove it from the others, then skip the next rule assuming the correctness of the initialized seed entities.
2. For each term $c \in C$, we compare its confidence score across all concepts, retain it in the concept where it has the highest score, and delete it from other concepts.

Finally, we employ a global optimization on every concept to filter noisy terms with sibling similarity scores below a predefined *threshold*. Let us note that the variables T, *sample-size*, *top_k_sg*, *top_k_entity*, *top_k_each_feature*,

top_k_type, *top_k_expand*, *top_k_per_concept*, and *threshold* are hyperparameters, that will be set in the experiments.

4.3 Location Recognition and Disambiguation

To identify geographical location entities in narrative texts, we utilize Named Entity Recognition (NER). NER is a popular data preprocessing task that seeks to locate and classify named entities mentioned in unstructured text into predefined categories such as person names, organizations, locations, time expressions, etc. In our study, we utilize several state-of-the-art (SOTA) NLP frameworks, including spaCy [14], FLAIR [1], and Stanza [22].

Geographical location disambiguation is crucial for visualizing migration routes. To pinpoint exact locations on a map, we need to link those recognized by NER to their accurate entries in a KB, a task known as Entity Linking (EL). EL is a common NLP task in practical applications, aiming to match textual entity mentions to KB entries like Wikipedia or Wikidata, serving as canonical entries. In our study, to link location entities to a KB, we propose using a recent and advanced tool called Bi-encoder Entity Linking Architecture, known as BELA [21]. Provided by Meta Open Source, BELA is the first transformer-based, end-to-end, one-pass, and multilingual EL model that efficiently identifies and links entities in texts, covering approximately 16 million entities and 97 languages. It utilizes a bi-encoder architecture that requires a single forward pass through a transformer for end-to-end linking of a passage, regardless of the number of entity mentions present. BELA has been trained on a Wikipedia dataset consisting of approximately 661 million samples, considering all the Wikipedia articles across 97 languages. For details regarding the training of BELA, we refer to Plekhanov et al. [21]. We choose BELA over other multilingual EL systems for several reasons. (1) It is computationally less expensive than other systems, such as [8], making it more practical for real-world applications. (2) For Entity Disambiguation (ED) tasks, it does not require a predefined candidate set for each mention, which is suitable for our case. (3) It disambiguates multiple entities in one pass, which is faster for disambiguation speed, while other systems [4, 8] require a mention encoding pass for each candidate.

BELA can be used for both End-to-End Entity Linking (EL) and Entity Disambiguation (ED) tasks. In our study, we utilize it for ED tasks, focusing on linking the location entities found in narrative texts to the knowledge base (KB) of Wikidata [31].

5 Experiments

We conduct the experiments for *MultiWidthExpan*, location detection and location disambiguation using the datasets presented in Table 1.

For *MultiWidthExpan* and location detection, the experiments are conducted on a computer equipped with Intel(R) Core(TM) i9-10900 CPU @ 2.80GHz with 62Gib of memory. For location disambiguation using BELA, the experiments are

conducted on a GPU machine equipped with Intel XEON Gold 6226R (16 cores @2.9GHz) and 4 NVIDIA Quadro RTX8000 (48GB graphics processors and 512 Gib of memory). The source codes and instructions for reproducibility are available here.

5.1 Experiments for MultiWidthExpan

We evaluate the quality of the expanded sets using *MultiWidthExpan* by comparing them to those annotated by experts. Given our small corpora, *MultiWidthExpan* is allowed to run until there are no additional quality terms for each concept. Table 2 reports the quantitative results using information retrieval metrics [5]. Hyperparameters in the first column are kept at their default values as in [27], except for *top_k_type* and *top_k_per_concept*, which are adjusted to adapt to our corpora. As we set *top_k_per_concept* = 50, the metrics in the last column are considered as P@50, R@50, and F1@50. However, after each iteration, since we set a predefined *threshold* ($t \in [0.7, 0.8]$) in the fifth column to filter out noisy terms, many terms fell below this threshold (i.e. no concept exceeds 50 terms in the sixth column, and will be truncated if it does). For the seed S, we discussed with HSS experts and conducted preliminary experiments to select the best entities (around 10% of the overall annotated entities) for each concept.

Despite having a small corpus and weak supervision of initialized terms, experiments show that *MultiWidthExpan* achieves desirable results by expanding many relevant terms from a list containing many noisy terms (see Table 1, only approximately 6% of terms are relevant for both types). Nevertheless, *MultiWidthExpan* expands some noisy terms, especially for *Accommodation*, which has the lowest precision and the highest recall for both types. This implies that while it expands many relevant terms, it also includes many noisy terms. Certain noisy terms share the same skip-pattern features with relevant terms in our corpora. For example, in English narratives, phrases like "I spent one month in one room" and "There is one kitchen" demonstrate that while "room" is a relevant term, "kitchen" is a noisy term. Similarly, in French narratives, phrases like "Je vis dans un foyer..." and "Je suis resté comme ça dans un jardin" show that while "foyer" is a relevant term, "jardin" is a noisy term. As shown in the fifth column, most of the red terms expanded in each concept have very close semantics with other terms inside their own concept but were not considered relevant based on the context of the narratives, as annotated manually by experts. For French narratives, *MultiWidthExpan* achieves higher F1-scores for all concepts compared to English narratives. One reason might be the number of narratives, since we have more narratives in French than in English, thus yielding more high-quality skip-pattern features. Since we rely on the CamemBERT MLM to generate the types of entities for French narratives, this suggests that this model can generate many pertinent types when provided with specific prompts, thus playing an important role in the expansion process.

However, as mentioned in [26], it is challenging to establish a perfect scoring method to obtain the ideal sets of entities, given the diversity and noisiness of unstructured text such as migration narrative texts in our study.

Table 2. Quantitative Results of MultiWidthExpan using Precision (P), Recall (R), and F1-score (F1). Annotated Set: ● missing term in Expanded Set. Expanded Set: ●, ● correct or incorrect expanded term respectively. ● initialized term (not include to calculate P, R, F1).

Type: hyperparameters	Concept	Annotated Set provided by Experts	S	t	Expanded Set generated by MultiWidthExpan	Metrics
English: *T=10, sample-size=20, top_k_sg=200, top_k_entity=30, top_k_each_feature=50, top_k_type=70, top_k_expand=10, top_k_per_concept=50*	**Accom.**	homes, home, house, houses, building, room, rooms, hotel, tent, tents, bed, apartment, shelter, floor, garage, hostel	home shelter	0.74	home, shelter, shops, parking, kitchen, houses, hotel, door, dormitories, bed, bathroom, floor, house, cabin, building, homes, doors, hostel, toilets, room, tent, tents, garage, shop, rooms, apartment, restaurant, accommodation	P: 53.84% R: 100% F1: 70%
	Means of Transport	autobus, boat, car, cars, foot, taxi, taxis, cab, bus, buses, minibus, van, vans, minivan, minivans, truck, trucks, tractor, train, trains, tram, ship, ferry, plane, airplane, flight, road, walk	boat train plane	0.72	boat, train, plane, airplane, trains, minivan, ship, flight, passengers, tractor, cab, taxi, vans, trucks, petrol, luggage, taxis, ride, car, cars, helicopters, buses, bus, ferry, tram, minivans, minibus, truck, passenger	P: 76.92% R: 80% F1: 78.43%
	Env.	jungle, jungles, forest, forests, trees, seaside, mountain, mountains, desert, hill, river, sea, valley, rocks, beach, coast, island, islands	jungle island	0.70	jungle, island, jungles, coast, region, mountain, hill, desert, sea, forests, mountains, southwest, valley, beach, area, forest, land, river, south, islands, seaside	P: 73.68% R: 87.50% F1: 80%
	Family or Friends	family, father, mother, brother, brothers, grandchildren, parents, dad, mum, mom, son, sons, sister, sisters, relatives, relative, cousin, cousins, uncle, wife, husband, kids, child, daughter, daughters, children, grandfather, grandmother, sister-in-law, friends, friend	mom relative daughter	0.75	daughter, mom, relative, child, grandfather, sons, daughters, brothers, relatives, friend, cousin, grandmother, uncle, cousins, sister, marriage, friends, nephew, son, boyfriend, dad, sisters, father, grandchildren, housewife, family, brother, elder, wife, mother, children, parents, husband	P: 83.34% R: 89.28% F1: 86.20%
French: *T=10, sample-size=20, top_k_sg=200, top_k_entity=30, top_k_each_feature=50, top_k_type=60, top_k_expand=10, top_k_per_concept=50*	**Accom.**	maison, appartement, chambre, chambres, maisonnette, hôtel, tentes, foyer, ferme, immeuble, domicile, tente, bâtiment, bâtisse, auberge, cabane, église, maisons, asile, salle	apparte-ment tente	0.72	appartement, tente, cabane, foyer, jardin, église, bâtisse, école, étage, bâtiment, asile, rue, village, immeuble, ferme, hôtel, maisonnette, maisons, ville, chambres, auberge, maison, chambre, domicile, salle	P: 73.91% R: 94.44% F1: 82.92%
	Means of Transport	bateau, camion, camions, pickup, marche, routes, pied, transport_commun, autobus, train, avion, convoi, véhicule, véhicules, taxi, taxis, métro, tram, bus, minibus, 4×4, autocar, voiture, voitures, vol, remorque, fourgonnette, route, autoroute, rail	avion taxi métro	0.76	avion, taxi, métro, fourgons, remorque, camions, rail, bateau, vélo, aéroport, camion, convoi, routes, tram, minibus, voiture, train, hélicoptères, bus, autobus, taxis, route, 4×4, véhicule, fourgonnette, autocar, voitures, chauffeur, hélicoptère, véhicules, autoroute	P: 78.57% R: 81.48% F1: 80%
	Env.	mer, forêt, forêts, montagne, désert, brousse, montagnes, collines, rivière, île, jungle, lac, broussaille, océan	forêt mer	0.80	forêt, mer, désert, rivière, forêts, île, océan, brousse, montagne, broussaille, terre, lac, montagnes, jungle, côte	P: 84.61% R: 91.66% F1: 88%
	Family or Friends	oncle, oncles, père, frère, grand-frère, ami, amis, grand-père, parent, parents, mère, famille, tante, sœur, sœurs, maman, frères, papa, cousin, cousins, tonton, fils, mari, frangin, cadet, amie, compatriote, compatriotes, compagnon, copine, copain, compagnons, fille, camarade, tuteur	parents frères cousin	0.75	parents, frères, cousin, famille, amie, ami, amis, copine, copain, homme, fils, compatriote, parent, frangin, oncle, oncles, maman, compatriotes, mère, femme, sœur, camarade, grand-frère, compagnon, jeune, cousins, tante, sœurs, frère, papa, mari, grand-père, tonton, pote, père, fille, cadet	P: 88.23% R: 93.75% F1: 90.91%

5.2 Experiments for Location Recognition

We evaluate spaCy, FLAIR, and Stanza by comparing their outputs to those annotated by experts. Table 3 reports the evaluations of these pretrained NER models on our corpora in terms of Precision, Recall, and F1-score.

According to the experiments, the pretrained models perform better on the English corpus than on the French corpus by approximately 10% in terms of F1-score. FLAIR achieves the highest F1-score for the English corpus, while spaCy achieves the highest F1-score for the French corpus. Interestingly, Stanza, which performs slightly lower on the English corpus, shows marginally better performance on the French corpus compared to FLAIR.

We observe that these pretrained NER models failed to recognize certain specific cities and small towns in various countries across different narratives in some parts of the texts, especially for French narratives, including *Bogovađa*, *Alipašino polje*, *Bazargan*, *Krnjača*, *Doğubayazıt*, *Tavo*, *Al-Shaykh Maskin*, *Gbèdjromèdé*, *Preševo*, *Kastanas*, *Orestiada*, *Polykastro*, *Horgoš*, *Sutukoba*, *Mansa Konko*, *Pakali Ba*, *Atatürk*, *Al-Salameh*, *Rakovica*, *Farmakonisi*, *Leros*, *Mytilini*, *Esendere*, *Sombor*, *El Jadida*, *Mohammédia*, *Alexandroúpoli*, *Loyané*, *Marvintsi*, *Evzonoi*, *Dojran*, *Ödemiş*, *Salmas*, *Zeytinburnu*, *Meissen*, and more.

Additionally, they also incorrectly predicted or classified some nationalities (e.g. *Nigérian(s)*, *Syrien(s)*, *Sénégalais*, *Malien(s)*, *Marocain(e)*, *Africain*, *Ivoirien(s)*, *Algérien(ne)*, *Guinéen(s)*, etc.), names of people (e.g. *Zireg*, *Lucie*, *Sami*, etc.), languages (e.g. *Pendjabi*, *Pashtu*, *Baloutchi*, etc.), and other entities (e.g. *Western Union*, *Schengen*, *Daesh*, etc.) as locations.

Table 3. Comparison of Pre-trained NER models for Location Recognition.

Model	NER-English			NER-French		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Spacy	0.9865	0.9619	0.9736	0.8651	0.9053	**0.8822**
FLAIR	0.9761	0.9756	**0.9755**	0.8615	0.9029	0.8785
Stanza	0.9842	0.9304	0.9561	0.8429	0.9335	0.8813

Table 4. Evaluation of Location Disambiguation using BELA.

Type	Ambiguous entities (countries or cities)	Accuracy
English	*Macedonia*, *Sombor*, *Banja*, *Dimitrovgrad*, *Skala*, *Rakovica*, *Kabal*, *Balochistan*, *Esendere*, *Skenderija*	0.9897
French	*Congo*, *Souba*, *Pogo*, *Siby*, *Farato*, *Tinzaouten* *Koloni*, *Konna*, *San*, *Bondo*, *Kalehe*, *Ondo*	0.9780

5.3 Experiments for Location Disambiguation

We evaluate BELA by comparing its output disambiguated locations to those disambiguated by experts. The results are shown in Table 4. The columns from left to right show the type of narratives, the ambiguous entities, and the accuracy.

Experiments show that BELA achieves high accuracy for both types, with performance on the English corpus slightly better than on the French corpus. However, there are some major ambiguous entities listed in the second column. For instance, in English narratives, BELA failed to recognize "Macedonia" as "North Macedonia" instead of the region in Greece. Another entity, "Sombor", a city in Serbia, was predicted as "Sumar", a city in Kermanshah Province, Iran. Similarly, in French narratives, BELA failed to distinguish "Congo" as "Democratic Republic of the Congo" instead of "Republic of the Congo". Another entity, "Souba", a village in the Ségou Region of Mali, was predicted as "Soba", a municipality in Spain. For other entities in the table, they show similar issues.

6 Map Visualization and Discussion with Experts

Figure 1(f) illustrates the process of visualizing the migration routes on a map. First, we employ a pretrained NER model to identify the locations mentioned in a narrative to obtain the tagged locations. Next, a pretrained disambiguation model based on the Wikidata KB is used to disambiguate those tagged locations to obtain their accurate Wikidata IDs. We then submit SPARQL queries to the Wikidata SPARQL endpoint using those Wikidata IDs to retrieve their Geonames IDs. Subsequently, we submit additional SPARQL queries to the Geonames Database [11] using those Geonames IDs to retrieve specific information about those locations, including longitudes and latitudes. Having obtained the longitudes and latitudes of such locations, we can visualize the migration routes on a map using OpenStreetMap [19]. The step-by-step details of this visualization process can be found here.

Our proposed framework has been presented, thoroughly analyzed and discussed with HSS researchers, experts, and the members of our project. It aids in accelerating the analysis of migration routes by automatically extracting relevant information and mapping out the migrants' journeys. Understanding the intricate nature and dynamics of migratory journeys requires examining numerous factors experienced and articulated by migrants during their movement. The initial phase of our framework allows us to derive a comprehensive set of terms from expert-defined concepts, capturing key terms conveyed in the narratives. This approach enables direct interrogation of the narratives by collecting terms associated with each concept as used by the migrants themselves. This understanding highlights various aspects such as the pivotal role of migrant families or friends, the nature of accommodations, modes of transportation, and the diverse environments traversed. Fig. 4 and Fig. 5 depict the examples of narrative texts with the highlighted key terms associated with five different concepts: *Location*, *Accommodation*, *Family or Friend*, *Environment*, and *Means of Transport*, in English and French, respectively.

In summary, the knowledge extracted is highly valuable for Humanities and Social Sciences researchers, providing essential assistance in analyzing migratory journeys. Visualizing the itineraries traversed by migrants on a map offers a powerful tool for enhancing observation and comprehension of migration routes. Fig. 6 demonstrates five primary *Trans-Saharan* routes traversing four African countries (Morocco, Mauritania, Algeria, and Mali), originating from Senegal, Guinea, Ivory Coast, Burkina Faso, and Nigeria.

LOC
Accommodation
Family Members
Environment
Means of transport

Asadabad is near the Kunar province . I heard that from other friends and from some people going on that way . I knew exactly that way will be easy to go to Europe : From Jalalabad , then Kabul , then Kandahar , then Quetta , then Quetta to the jungle , and from jungle to Iran until Istanbul . I know exactly that I needed to go to that place first , then that place , and then that place . I ask someone where I can sleep , and where I cannot sleep . The night is coming , I think about finding somewhere to sleep . Sometimes in the jungle , give money to someone to have a bed . The taxi driver knew about it . I gave money to him and he told me . In some jungles , my friends passed this way . Those guys tell their stories . Some of them were shot . I am scared of Asadabad . Here it 's too hard , but my life is more dangerous there . I did n't see any hot time because everyday everybody is like Serbian , no fight , nothing just safe and everyone is good , that 's a very safe place . Many Afghans are in Aksaray . No one speaks English , this is a big problem , but Rehan is a good learner . In Turkey , the taxi driver told us how to go . Many guys are waiting . He said he will find something that will help us to cross . In Greece , it 's very dangerous because there is the sea . You can be only 20 people , but they put 60 people on one boat . So , there is a big danger . That way is near but too dangerous . That way is far but less dangerous . There was a guy that the taxi driver knows , and he showed the way . But the person told us : you go on that way and on that way . When we reached there , there were taxis to go to Sofia . That taxi driver took us to a place where we can sleep . But the biggest problem is we do n't know the Bulgarian language . They took our money ; they asked us for hundred dollars to go to Serbia . The taxi took us to the jungle , and again , the same guy we met , he shows us the way : you should go on this way . We did n't know anything about Bulgaria and we cannot go out of the house . We were afraid of the police and the fingerprints . That 's why many of us are going through Bulgaria . It was n't very expensive . Then we move to the jungle , we stayed all day in the jungle . There are houses , so if someone sees you , they will call the police . So better if you stay in the jungle . Then you cross the border . You need to reach the main road to Dimitrovgrad . The guy explained us before . After that , we travelled there by bus , thanks to NGOs . From Istanbul , there are only two ways to go to Europe . One is the sea through Greece , and one is Bulgaria . But people told us to be aware , to be not arrested by police that will take fingerprints and then will deport us from Europe to Bulgaria . We know that if you cross Bulgaria to enter Serbia , then , no problem . When we reached Dimitrovgrad , UNHCR and Red Cross gave us food , explained where we are , and gave us clothes , blankets , and everything , they arrange a ticket for us to go to Belgrade . Now , we are waiting for friends who are still in Bulgaria . We only know about Greece , that we needed to cross the sea . We knew from Greece that we need to go to Serbia and then Hungary . But , when we were in Istanbul , they said you could not go to Hungary , it 's a danger for life . So , we decide to go on that way . It 's a little bit expensive but it 's the safest way . Greece is less expensive but too dangerous . Afghanistan , fear for life ; Turkey is the best place I 've always been ; Bulgaria is the most frightening place that no one can forget , there were too many problems , Bulgarian police and Bulgarian people all treated us very bad ; and Serbia is better than Turkey because of the people , the friends we got here , all the people we met here .

Fig. 4. Example of a narrative text in English.

LOC
Hébergement
Membres de famille
Environnement
Moyens de transport

Bobo Dioulasso, juillet 2009, 06 septembre 2009, j'ai pris cette route parce que j'ai volé l'argent de mon frère pour fuir. C'est quand je suis arrivé à Koutiala au Mali on m'a parlé de cette route. C'est pour cette raison que j'ai tenté de venir ici. C'est des amis que j'ai rencontrés au Mali qui m'ont informé de cette route. Je suis passé par Koutiala, Bamako, Gao, Bordj Badji Mokhtar, Riganne et Bordj Badji Mokhtar où je suis resté. Je n'avais aucune connaissance sur les routes ni la difficulté. J'ai voyagé pour rentrer à Koutiala parce que j'ai peur que mon grand-frère me frappe. Je voulais me retourner à Bobo mais j'avais déjà dépensé beaucoup donc j'ai finalement opté pour le voyage sur l'Europe. Je suis resté avec une famille malienne sous prétexte que je suis orphelin. Après que les amis m'ont parlé de la route j'ai pris un minibus pour aller à Bamako. Après, je suis parti à Gao. J'ai payé 75.000 FCFA (environ 110 euros) pour passer en Algérie. J'ai emprunté le pickup avec une dizaine de migrants. On était très serré dans le pickup. J'étais fatigué et tout le monde criait et demandait au chauffeur d'arrêter le pickup. C'est un très difficile voyage. On a pris deux jours pour arriver à Bordj Badji Mokhtar. Le pickup nous a laissés à environ deux kilomètres du poste frontière. On a marché pour aller au poste. Les policiers ont pris tous les passeports pour le contrôle. J'étais de ceux qui n'ont pas pu traverser la frontière avec la faute de passeport. Alors le lendemain j'ai payé de l'argent à un démarcheur et j'ai finalement traversé la frontière. J'avais besoin de passeport pour continuer le parcours alors j'en ai payé un. J'ai pris le transport commun pour aller à Riganne. Je suis resté là-bas avec des maliens. Je me suis enfin mis à travailler parce que je commençais déjà à en manquer. Un jour, des maliens voulaient continuer le chemin et j'ai fait le chemin avec eux. On a emprunté un autobus. Comme on était au nombre de six subsahariens la police nous a contrôlé beaucoup de fois. Depuis que je suis arrivé ici je suis maintenant très décidé à continuer mon chemin jusqu'en Europe parce qu'il n'y a rien en Afrique. Je suis dans le foyer malien (appartement de trois pièces). Je travaille dans la construction et je vis de cet argent. Les gens sont très sympas et je n'ai pas de dérangements. Je n'ai pas grand contact avec les algériens. Je n'ai pas de contact avec mes parents. Ici la police rôde dans les environs mais ne font pas de rafles. C'est au Tchad (lieu où les migrants s'arrêtent pour chercher du travail que la police vient prendre les gens. Je ne sais vraiment pas ce qu'il faut penser de l'avenir.

Fig. 5. Example of a narrative text in French.

Fig. 6. Itineraries of migration routes from the Trans-Saharan region.

7 Conclusion and Future Works

In this paper, we have presented a text mining approach to leverage the collection of migration narratives in English and French. We first collaborated with HSS experts to annotate the essential concepts, their related terms, and the locations mentioned in the narratives. We then adapted a set expansion algorithm to extract related terms embedded in these narratives in a weakly supervised manner using a small set of annotated terms. We utilized existing NER models to identify locations crossed by migrants, followed by a pretrained disambiguation model to precisely locate them on a map. Experiments were conducted by comparing the

output generated by the proposed algorithm and models to those annotated by experts. Finally, we design a tool to visualize the itineraries of migration routes on a map. Insightful discussions with HSS experts concluded that our framework aids their analyses by automatically extracting relevant terms and drawing the itineraries of migrants, providing a deep understanding of migration phenomena.

In the future, we plan to collect more interview texts while considering additional sensitive concepts, such as control and human trafficking. Addressing these non-neutral or politically charged concepts may require a larger dataset of migrant narratives to train and generate specific embeddings, thereby minimizing noisy terms during the expansion process. Additionally, we aim to enhance the performance of NER models, particularly for French, by fine-tuning pretrained models on French narratives and reassessing their effectiveness. Our next goal is to use the extracted terms from each concept to explore their relationships and convert the unstructured text into a machine-readable format by constructing a migration ontology in a (semi-)automatic manner with less human intervention. Finally, exploring the capabilities of more complex and sophisticated models like large language models on migration narratives is a promising research direction.

Acknowledgements

Many thanks to the reviewers for their insightful comments and suggestions. This work has benefited from the support of the region Hauts-de-France and ANR HYCI Project (ANR-22-CE55-0010) of the French National Research Agency.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL. pp. 54–59 (2019)
2. Bacon, L.: La fabrique du parcours migratoire sur la route des Balkans. Co-construction des récits et écritures (carto)graphiques. Ph.D. thesis (2022)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the ACL **5** (2017)
4. Botha, J.A., Shan, Z., Gillick, D.: Entity Linking in 100 Languages. In: EMNLP. pp. 7833–7845 (2020)
5. Carterette, B., Voorhees, E.: Overview of Information Retrieval Evaluation, pp. 69–85 (01 2011)
6. Chen, Z., Cafarella, M., Jagadish, H.V.: Long-tail Vocabulary Dictionary Extraction from the Web. In: WSDM. p. 625–634 (2016)
7. Craswell, N.: Mean Reciprocal Rank. Springer US (2009)
8. De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., Petroni, F.: Multilingual Autoregressive Entity Linking. Transactions of the ACL **10** (2022)
9. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition of Multi-word Terms: The C-value/ NC-value Method. vol. 3, pp. 115–130 (2000)
10. Gantz, J., Reinsel, D.: Extracting value from chaos. IDC iview (2011)
11. GeoNames: Geonames. `http://geonames.org/`, retrieved June 17, 2009

12. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning Word Vectors for 157 Languages. In: LREC (2018)
13. He, Y., Xin, D.: SEISA: set expansion by iterative similarity aggregation. In: The Web Conference (2011)
14. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020)
15. Hussain, M.N., Bandeli, K.K., Al-khateeb, S., Agarwal, N.: Analyzing shift in narratives regarding migrants in Europe via blogosphere (2018)
16. Ioffe, S.: Improved Consistent Sampling, Weighted Minhash and L1 Sketching. In: ICDM. pp. 246–255 (2010)
17. Macanovic, A.: Text mining for social science – The state and the future of computational text analysis in sociology. Soc. Sci. Res. (2022)
18. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: CamemBERT: a Tasty French Language Model. In: ACL (2020)
19. Nelson, A., de Sherbinin, A., Pozzi, F.: Towards development of a high quality public domain global roads database. Data Sci. J. (2006)
20. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language Models as Knowledge Bases? In: EMNLP-IJCNLP (2019)
21. Plekhanov, M., Kassner, N., Popat, K., Martin, L., Merello, S., Kozlovskii, B., Dreyer, F.A., Cancedda, N.: Multilingual End to End Entity Linking (2023)
22. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: ACL: System Demonstrations (2020)
23. Robin, N.: Migrations, observatoire et droit. complexité du système migratoire ouest-africain. migrants et normes juridiques. In: HdR. Univ. de Poitiers (2014)
24. Rong, X., Chen, Z., Mei, Q., Adar, E.: EgoSet: Exploiting Word Ego-networks and User-generated Ontology for Multifaceted Set Expansion. In: WSDM (2016)
25. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
26. Shen, J., Wu, Z., Lei, D., Shang, J., Ren, X., Han, J.: SetExpan: Corpus-Based Set Expansion via Context Feature Selection and Rank Ensemble (2019)
27. Shen, J., Wu, Z., Lei, D., Zhang, C., Ren, X., Vanni, M.T., Sadler, B.M., Han, J.: HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion (2019)
28. Shi, S., Zhang, H., Yuan, X., Wen, J.R.: Corpus-based Semantic Class Mining: Distributional vs Pattern-Based approaches. In: COLING (2010)
29. Taylor, W.L.: "Cloze Procedure": A New Tool for Measuring Readability. Journalism Quarterly **30**(4), 415–433 (1953)
30. Tong, S., Dean, J.: System and methods for automatically creating lists. US Patent 7,350,187. (2008)
31. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
32. Wang, C., Chakrabarti, K., He, Y., Ganjam, K., Chen, Z., Bernstein, P.A.: Concept Expansion Using Web Tables. In: WWW (2015)
33. Wang, R.C., Cohen, W.W.: Language-Independent Set Expansion of Named Entities Using the Web. In: ICDM. p. 342–350 (2007)
34. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: A Probabilistic Taxonomy for Text Understanding. In: SIGMOD. p. 481–492 (2012)
35. Öztürk, N., Ayvaz, S.: Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. Telemat. Inform. (2018)