GeoLingIt at EVALITA 2023: Overview of the Geolocation of Linguistic Variation in Italy Task

Alan Ramponi[1,*], Camilla Casula[1,2]

[1]Fondazione Bruno Kessler (FBK), Digital Humanities Unit – Trento, Italy
[2]University of Trento, Department of Information Engineering and Computer Science – Trento, Italy

Abstract

GeoLingIt is the first shared task on geolocation of linguistic variation in Italy from social media posts comprising content in language varieties other than standard Italian (i.e., regional Italian, and languages and dialects of Italy). The task is articulated into two subtasks of increasing complexity for which only textual content is allowed: i) *coarse-grained geolocation*, aiming at predicting the region in which the variety expressed in the post is spoken, and ii) *fine-grained geolocation*, aiming at predicting its exact coordinates. Both tasks can be either at the country level (*standard track*) or restricted to a linguistic area of choice (*special track*). GeoLingIt has attracted wide interest at the Evalita 2023 evaluation campaign with 37 registrations and 35 submitted runs. In this paper, we present the task and data, the evaluation criteria, the participants' results, an analysis of their approaches, and the main insights from the shared task.

Keywords

Natural language processing, computational sociolinguistics, linguistic variation, linguistic diversity

1. Introduction

Italy is characterized by an astonishing linguistic diversity that makes it a unique landscape in Europe [1]. Besides standard Italian, a large number of local languages, their dialects, and regional varieties of standard Italian (i.e., regional Italian) are spoken across the country [2]. While Italian is employed in all formal settings in its standard form, in informal situations it is natural to observe Italian speakers to use (even unwittingly) regional forms of Italian (e.g., *guaglione*, *toso*, and *caruso* for "young man", typically in Campania, Lombardy-Veneto, and Sicily areas, respectively), or to code-switch their local language varieties with the national language.

Local languages and their dialects evolved from Vulgar Latin like Italian, and they mostly have no established orthography insofar as they are primarily used in spoken settings. On the other hand, regional forms of Italian derive from a geographical differentiation of Italian due to influences by the former [3], are largely used in both oral and written informal contexts, and typically follow Italian spelling conventions. When it comes to user-generated texts on social media, which are informal and feature linguistic patterns from spoken language [4, 5], we observe that not only regional Italian is naturally present, but also local language varieties of Italy are employed, albeit at various degrees. This can be attributed to their rediscovery as "additional expressive resources" [6], especially by the youngest generations. User-generated texts comprising language varieties other than standard Italian open opportunities for the study of linguistic variation in Italy, and can ultimately help in enriching and complement linguistic atlases.

In this paper, we present GeoLingIt, the first shared task on geolocation of linguistic variation in Italy from social media posts from Twitter containing content other than standard Italian. GeoLingIt has been organized as part of the Evalita 2023 evaluation campaign [7], and relies on DiatopIt [8], a corpus of geolocated tweets exhibiting regional Italian use, code-switching between Italian and local language varieties, or fully written in the latter. Compared to previous geolocation shared tasks at international venues [9, 10, 11], GeoLingIt is focused on Italy and tailored to variation across language varieties, and it thus minimizes the effect of spurious, highly-localized lexical items (e.g., mentions of events, places, or tourist attractions) on prediction of linguistic areas. In the following, we present details on GeoLingIt, the results obtained by participant teams, and the main insights from the shared task.

2. Task description

The GeoLingIt shared task deals with the geolocation of linguistic variation in Italy from Twitter posts comprising content in language varieties other than standard Italian (i.e., regional Italian, and languages and dialects of Italy). It aims to advance the study of linguistic variation in Italy, provide means to complement qualitative-driven linguistic atlases, and sensitize the community on the rich linguistic landscape of the country.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
alramponi@fbk.eu (A. Ramponi); ccasula@fbk.eu (C. Casula)

2.1. Standard and special tracks

GeoLingIt is organized into two tracks. In the *standard track*, the focus of the task is at the country level (i.e., comprising all language varieties of Italy), whereas in the *special track*, the task is restricted to a linguistic area chosen by participants[1] (e.g., the *Gallo-Italic area*, including language varieties spoken in Piedmont, Lombardy, Liguria, and Emilia-Romagna regions) to favor the emergence of microvariation insights. For both tracks, two subtasks of increasing complexity are possible: *coarse-grained geolocation* (Section 2.2) and *fine-grained geolocation* (Section 2.3).

2.2. Subtask A: Coarse-grained geolocation

Given the text of a tweet exhibiting regional Italian features or (partially or fully) written in local languages and dialects of Italy, predict the administrative region in which the variety expressed in the post is spoken. This is a classification task, i.e., one among n regions of Italy has to be predicted. In the case of the *standard track*, this matches all regions of Italy[2] ($n = 20$), whereas in the *special track*, it corresponds to the subset of regions k of the linguistic area under consideration ($n = k$). This subtask is applicable for the *special track* if $k \geq 2$ regions are represented in the chosen area.

2.3. Subtask B: Fine-grained geolocation

Given the text of a tweet exhibiting regional Italian features or (partially or fully) written in local languages and dialects of Italy, predict the location, in terms of longitude and latitude coordinates, in which the variety expressed in the post is spoken. This is a double regression task, i.e., a pair of real-valued numbers has to be predicted. The difference between *standard* and *special* tracks is here the extent of the area being considered. This subtask overcomes the simplification of *coarse-grained geolocation* (Section 2.2), aiming to uncover fine-grained linguistic variation. Indeed, language varieties of Italy lie on a continuum and often cross administrative region borders.

3. Data

GeoLingIt is based on DiatopIt [8], a corpus of social media posts from Twitter specifically focused on language variation in Italy. All tweets have associated geolocation information and region labels, and have been sampled to contain either regional Italian usage or content in local language varieties of Italy. A multi-stage data collection process has been followed based on data-driven out-of-vocabulary tokens (from posts over a period of 2 years) which have been curated manually. Under-represented areas from the resulting posts have been then augmented by employing the lexical artifacts package [13]. The corpus consists of 15,039 posts from a 2-year time frame (from `2020-07-01` to `2022-06-30`) to minimize period-related biases. For more details, we refer the reader to Ramponi and Casula (2023) [8].

Data splits During the development stage, participant teams are provided with the original training and development splits of DiatopIt. These splits consist of 13,669 and 552 examples, respectively. While the training set comprises content from all over the country, the development set contains data from 13 out of 20 regions.[3] Teams are allowed to use alternative splits and even augment the dataset at their will, with the only constraint to not use external Twitter data since some tweets can be part of the test set. The (unlabeled) test set is then released during the evaluation window for allowing teams to submit their predictions, and comprises 818 examples from the same regions in the development set plus examples from $1 \leq j \leq 7$ additional regions unknown to participants during both development and evaluation stages. At the end of the evaluation window, the $j = 4$ additional regions in the test set have been communicated to participants.[4] Splits match the original data partitions of DiatopIt; we thus refer the reader to Ramponi and Casula (2023) [8] for details on statistics and distribution.

Data format The corpus splits are in the form of `tsv` files, i.e., a tab-separated format, with an example per line and the first line as header. Each example has `id` and `text` columns. For the *coarse-grained geolocation* subtask, data files additionally include a `region` column, whereas data files for the *fine-grained geolocation* subtask include `latitude` and `longitude` columns. As a result, the instances in both the subtasks are the same, and differ according to the label column(s). The content of such columns is described below:

- **id**: a unique identifier, different from the original tweet identifier to preserve user's anonymity;
- **text**: the text of the tweet, with anonymized user mentions, email addresses, URLs, and location strings deriving from cross-platform posting;

[1] Participants have been provided with the renowned linguistic map by Pellegrini (1977) [12] to encourage linguistically-grounded proposals, and requests have been approved based on motivation and relevance of the area from a linguistics perspective.

[2] These are: Abruzzo, Aosta Valley, Apulia, Basilicata, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lazio, Liguria, Lombardy, Marche, Molise, Piedmont, Sardinia, Sicily, Tuscany, Trentino-Alto Adige, Umbria, and Veneto.

[3] Regions in the development set: Apulia, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lazio, Liguria, Lombardy, Piedmont, Sardinia, Sicily, Tuscany, and Veneto.

[4] Regions in the test set: the regions in the development set plus Abruzzo, Marche, Trentino-Alto Adige, and Umbria.

Table 1
Examples from the DiatopIt corpus [8] used for the GeoLingIt shared task. Columns in *yellow* and *gray* are included in subtask A and B files, respectively. *Note: IDs have been changed and texts have been slightly redacted to preserve users' anonymity.*

id	text	region	latitude	longitude
1	chiov' tutt a jurnat', ce serv' o mbrell'	Campania	40.92589400	14.12023790
2	ho così sonno che me bala l'oeucc	Lombardia	45.46134530	9.15933655
3	da caruso anche io ci andavo spesso!	Sicilia	37.46007165	15.03084825

- **region**: the administrative region from where the tweet has been posted, in a string format (for *subtask A* only);
- **latitude**: a real-valued number representing the latitude coordinate from where the tweet has been posted (for *subtask B* only);
- **longitude**: a real-valued number representing the longitude coordinate from where the tweet has been posted (for *subtask B* only).

The `latitude` and `longitude` coordinates are computed by taking the central point from the 4-point bounding box of city areas as provided by the Twitter APIs. Note that coordinates do not correspond to specific places within cities, but instead represent cities as a whole (i.e., posts originated within the same city have the same coordinates). We provide examples from the corpus in Table 1.

4. Evaluation

During the evaluation phase, participant teams are allowed to submit up to 3 runs (i.e., predictions on the unlabeled test set) for each track and subtask. In all the setups, only textual content can be used. We here present the metrics used for assessing the performance of runs (Section 4.1) and the baselines we provide (Section 4.2).

4.1. Metrics

Due to the different nature of *coarse-grained geolocation* and *fine-grained geolocation*, we employ different evaluation metrics for the subtasks. Subtask-specific metrics are the same for both *standard* and *special* tracks.

Subtask A The submitted runs are evaluated using macro-averaged precision, recall, and F_1 score on the n regions of Italy under consideration. For the *standard track*, this matches all the administrative regions in the test set ($n = 17$, cf. Section 3, "Data splits"), whereas for the *special track*, this corresponds to the k regions in the chosen linguistic area that are also represented in the test set ($n = k$, cf. Section 2.2). Runs are ranked by macro F_1 score and presented in separate rankings (i.e., one for the *standard track*, and one for each chosen subset of administrative regions in the *special track*).

Subtask B Since a level of detail to the centimeter is unnecessary to the study of linguistic variation, runs are evaluated using the average distance in kilometers (km) of predicted coordinates from gold coordinates (the lower the better) on either the whole test set (for the *standard track*) or the subset representing the linguistic area chosen by participants (for the *special track*). We employ the Haversine formula as implemented in the `haversine` (v2.8.0) package[5] for computing distance. Runs are presented in separate rankings (i.e., one for the *standard track*, and one for each area in the *special track*) and ordered by increasing average distance in kilometers.

4.2. Baselines

We use the same baselines for both tracks. For subtask A, we provide a most frequent baseline and a logistic regression baseline. For subtask B, we provide a centroid baseline and a k-nearest neighbors baseline.

Most frequent A baseline that always guesses the most frequent administrative region in the training set (i.e., Lazio) for all test set instances.

Logistic regression A machine learning classifier with default `scikit-learn` (v1.2.2)[6] hyperparameters that employs count vectorizer with unigrams for feature extraction and operates on original text casing.

Centroid A baseline that computes the center point (in terms of latitude and longitude) from the training set and predicts it for all test instances.

k-nearest neighbors (kNN) A machine learning regressor with default `scikit-learn` hyperparameters that employs count vectorizer with unigrams for feature extraction and operates on original text casing.

5. Participants and results

A total of 35 runs have been submitted to the GeoLingIt shared task: 26 runs (6 teams) for the *standard track* and

[5] `haversine` package: https://github.com/mapado/haversine
[6] `scikit-learn` library: https://scikit-learn.org

9 runs (2 teams) for the *special track*. Specifically, for the *standard track* we received 14 runs (5 teams) for subtask A and 12 runs (5 teams) for subtask B, whereas for the *special track* 6 runs (2 teams) have been submitted for subtask A (i.e., *Tuscany-Lazio area* and *Gallo-Italic area*) and 3 runs (1 team) have been tailored at subtask B (i.e., *Gallo-Italic area*). Overall, GeoLingIt has been one of the most participated shared tasks at Evalita 2023 [7] and attracted interest of heterogeneously composed teams with up to 7 individuals, from master students to senior academic researchers.

5.1. Overview of participant teams

In the following, we provide a summary of the approaches employed by participant teams. We refer the reader to their description papers for additional details.[7]

baρtti [14] The team participated in both subtasks for the *standard track*. For subtask A, they experimented with multi-task learning, a transformer-based and logistic regression model ensemble, and contrastive pre-training of a BERT-based Italian model on augmented subtask data. Augmentation uses a vocabulary built from online sources to create examples by randomly substituting words with lexical items from varieties spoken in the same or different regions. For subtask B, they leveraged data from both subtasks in a multi-task setting using either a BERT-based Italian model or the model that underwent continuous pre-training in subtask A, also testing a rectification module to adjust predictions outside land to the closest point within Italy's boundaries.

DANTE [15] The team focused on further pre-training BERT-based Italian language models and participated in both subtask A and B for the *standard track*. Specifically, they experimented with two multi-task pre-training setups, namely task-specific learning and joint learning, with dialect and token classification objectives, using texts collected from external sources. Fine-tuning is then done in a single task setup on relevant subtask data. In both subtasks, they also proposed ensembles of their best-performing models.

extremITA [16] The team proposed two *one-for-all* models, designed to tackle all the challenges at Evalita 2023. The first model is based on the IT5 encoder-decoder architecture, whereas the second one is an instruction-tuned model built upon LLaMA. For fine-tuning, they used data from all Evalita 2023 challenges and encoded the tasks as prompts. The team submitted a run for model for both subtasks of the *standard track*.

galliz [17] The team proposed a hybrid approach for subtask A, and participated in both the *standard track* and *special track*. Specifically, they combined the predictions given by *i)* an English pre-trained BERT classifier, previously fine-tuned on augmented GeoLingIt training data, and *ii)* a dictionary-based algorithm derived from external lexical sources. They then tested different hyperparameter setups. As regards data augmentation, the team fine-tuned an Italian word embedding model on the training set, and leveraged word vector similarities to create new training examples by substituting a single word per post with a close word in the embedding space.

Salogni [18] The team tested different transformer-based models pre-trained on Italian texts, with a set of hyperparameter settings (e.g., hidden layers, activation functions). They submitted a single run for the *standard track*, subtask B, based on a UmBERTo language model.

SCG The team participated to both tracks and experimented with logistic regression and support vector machines for subtask A, and linear regression and kNN regression for subtask B.[8] They did not submit a report and we are thus unable to discuss further their approach.

5.2. Results

In this section, we summarize the results of participant teams in both subtask A and B for the *standard track* (Section 5.2.1) and the *special track* (Section 5.2.2).

5.2.1. Standard track

We present the results divided by subtask below.

Subtask A: Coarse-grained geolocation In Table 2, we report the results on the test set for all runs submitted by teams participating in subtask A, ranked by macro F_1.

All runs by the *DANTE* team obtained the best results in the subtask, with improvements ranging from 5.52 to 10.10 macro F_1 points compared to the best run by the team that ranked second (*galliz*). The best-performing system by *DANTE* (run 3) is an ensemble of transformer-based classifiers originally pre-trained on Italian texts, which have been further pre-trained in a multi-task fashion on external data from Dialettando[9] and Wikipedia editions for local language varieties of Italy with region-centric objectives. The best submission by *galliz* (run 1) is an equally-weighted ensemble of a dictionary-based algorithm (based on Dialettando and GeoLingIt) and an English BERT model fine-tuned on augmented subtask A data, whereas the best run for *baρtti* (run 2) relies on a

[7] Indeed, we do not include the specific model versions and hyperparameter choices of participants' systems due to space constraints.

[8] We thank the SCG team for providing us with this information.

[9] "Dialettando" website: https://www.dialettando.com

Table 2
Results on the test set for the *standard track*, subtask A. Baselines are italicized and highlighted in *yellow*.

	Team	Run	P	R	F_1
1	DANTE	(3)	79.46	63.75	66.30
2	DANTE	(2)	66.98	62.65	63.93
3	DANTE	(1)	65.18	60.09	61.72
4	galliz	(1)	82.94	52.25	56.20
5	baρtti	(2)	67.97	51.62	53.18
6	galliz	(3)	74.58	49.49	52.08
7	baρtti	(3)	52.93	51.75	51.74
8	baρtti	(1)	56.05	51.68	51.72
9	galliz	(2)	68.98	45.36	47.74
	Log. reg.		*62.19*	*42.43*	*46.11*
10	extremITA	(1)	72.14	38.84	39.99
11	extremITA	(2)	65.03	37.62	38.18
12	SCG	(2)	12.92	9.82	9.28
13	SCG	(1)	10.15	9.97	9.04
14	SCG	(3)	10.42	6.60	7.85
	Most freq.		*4.47*	*21.15*	*7.38*

Table 3
Results on the test set for the *standard track*, subtask B. Baselines are italicized and highlighted in *yellow*.

	Team	Run	Avg dist (km)
1	baρtti	(3)	97.74
2	baρtti	(1)	98.79
3	DANTE	(3)	110.35
4	DANTE	(2)	112.58
5	DANTE	(1)	114.00
6	baρtti	(2)	120.02
7	extremITA	(1)	126.10
8	Salogni	(1)	128.19
9	extremITA	(2)	145.15
	kNN		*263.35*
10	SCG	(1)	280.99
	Centroid		*281.04*
11	SCG	(2)	281.20
12	SCG	(3)	289.91

transformer-based classifier, pre-trained on Italian texts, that has been further pre-trained in a contrastive learning fashion with subtask A data, preemptively augmented with a word substitution approach based on a vocabulary derived from Dialettando and Wikipedia content. While all teams outperformed the most frequent baseline, all runs by *extremITA* and *SCG* teams achieved worse results than the logistic regression baseline.

From a closer look, we observe that F_1 scores obtained by participants' runs greatly differ across regions (Figure 1). Campania, Lazio, Sardinia, Sicily, and Veneto are the easiest to classify. As expected, Abruzzo, Marche, Trentino-Alto Adige, and Umbria are instead among the regions with the lowest scores on average. This is mainly because posts from those regions have been excluded on purpose from the development set, and only few tweets are available in the training set, making traditional learning and tuning challenging. As a result, most instances from those regions are typically classified as neighboring regions in which similar varieties are spoken (e.g., posts comprising content in *Trentino* as spoken in the province of Trento – whose linguistic features exhibit traits of continuity between Lombard and Venetian [12] – are classified as Lombardy and Veneto, respectively).

Moreover, Friuli-Venezia Giulia and Apulia exhibit low scores on average across runs despite being represented in all data splits. The reason behind this has to be researched in linguistics rather than computation. Besides Friulian, Slovene and German varieties, in Friuli-Venezia Giulia varieties of Venetian are also spoken (e.g., the *Triestino* variety) [12], and thus posts comprising the latter are easily misclassified with the region in which Venetian is predominantly used (i.e., Veneto). On the other hand, *Salentino* varieties as spoken in the southern part of Apulia are part of the extreme southern varieties group [12], which also includes Sicilian, and thus make a large fraction of posts from Apulia to be misclassified as Sicily [8]. Besides the limitations of subtask A, this highlights that NLP should eventually go beyond "raw modeling" and start considering again linguistics as its foundation.

Subtask B: Fine-grained geolocation Test set results for all submitted runs in subtask B are reported in Table 3.

All teams except *SCG* outperformed both the baselines. The *baρtti* team obtained the best results with two out of three submissions (i.e., run 3 and 1). Their best run relies on multi-task learning on subtask A and B data, and uses geography-informed postprocessing to ensure that predictions fall inside the country borders. *DANTE*'s runs adopted similar methods to those employed in subtask A with separate layers for regression, ranking third with a model ensemble (run 3). *Salogni*'s run is based on UmBERTo fine-tuning, whereas the best run by *extremITA* is based on IT5 trained to generate region labels.

By looking at predictions by models that outperformed both baselines, we observe that, on average, errors range from 0.89 km to 668.11 km, with a median of 58.77 km. Errors are typically due to lexical items that are highly represented in other locations, e.g., posts with "*ghe* mel" (en: "of course", *Parmigiano* variety) fall in the Treviso area (Veneto) instead of the Parma area (Emilia-Romagna).

5.2.2. Special track

We present the results divided by subtask below.

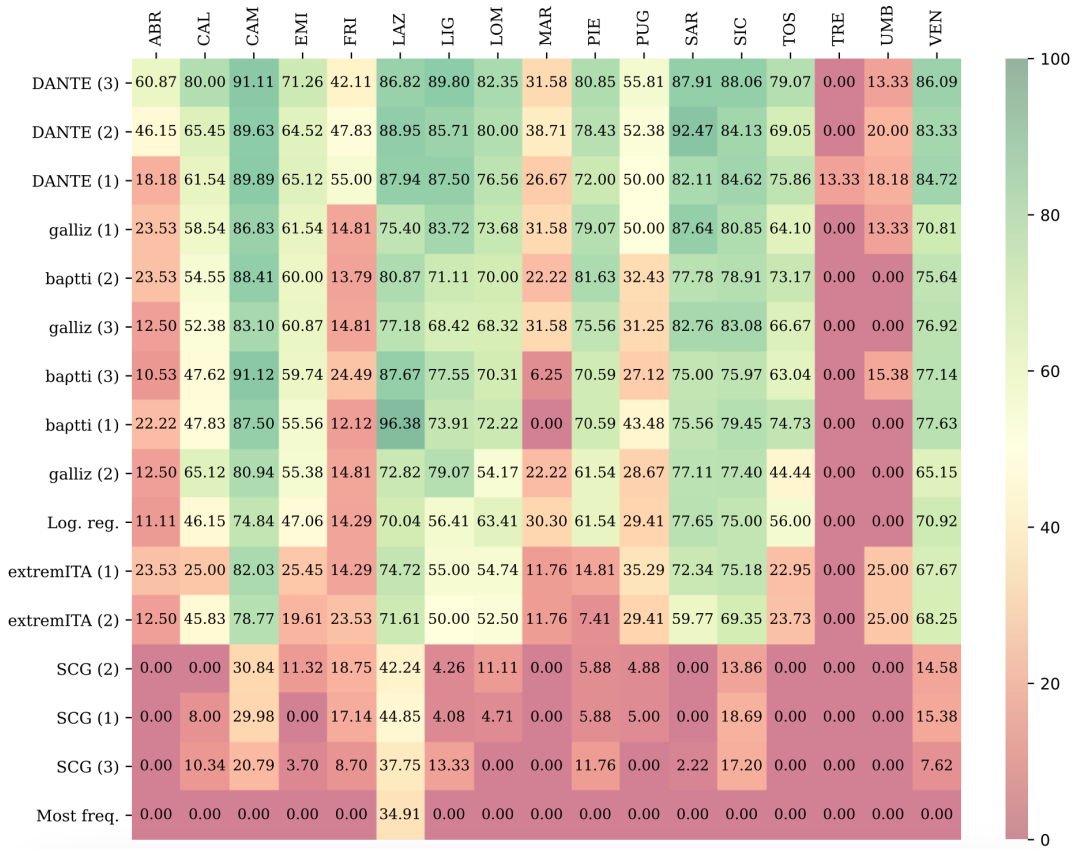

Figure 1: Results divided by region for the *standard track*, subtask A, in terms of F_1 score. Teams (with run numbers within parentheses) are on the rows, and test set regions (presented with their first three letters in Italian) are on the columns.

Subtask A: Coarse-grained geolocation Official results on the test set for the areas chosen by participant teams in subtask A (i.e., the *Tuscany-Lazio* area and the *Gallo-Italic* area) are summarized in Table 4.

As regards the *Tuscany-Lazio* area, the best run by the *galliz* team (run 3) achieved an improvement over the logistic regression baseline of 11.67 points in macro F_1 score. They employed a similar solution as the one for the *standard track*, additionally leveraging lexicons relevant to the linguistic area under consideration (i.e., lemmas from the *Vocabolario del Fiorentino Contemporaneo*[10] and a word list for the Romanesco dialect)[11] giving more weight to the BERT-based model. This confirms the usefulness of using region-specific linguistic materials in the task. For the *Gallo-Italic* area, all runs by the *SCG* team are between the two baselines we provided, but we are unfortunately unable to provide insights on their results.

[10]"Vocabolario del Fiorentino Contemporaneo" website: https://www.vocabolariofiorentino.it

[11]Romanesco word list from "The Roman Post" website: https://www.theromanpost.com/2016/06/dizionario-dialetto-romanesco

Table 4
Results on the test set for the *special track*, subtask A, divided by area. Baselines are italicized and highlighted in *yellow*.

	Team	Run	P	R	F_1
	Tuscany-Lazio area				
1	galliz	(3)	81.25	83.32	82.20
2	galliz	(1)	72.43	80.42	73.40
3	galliz	(2)	72.43	80.42	73.40
	Log. reg.		*91.79*	*66.67*	*70.53*
	Most freq.		*38.62*	*50.00*	*43.58*
	Gallo-Italic area				
	Log. reg.		*76.24*	*62.49*	*66.32*
1	SCG	(1)	30.86	31.36	29.14
2	SCG	(2)	29.42	29.03	26.66
3	SCG	(3)	25.30	26.25	22.78
	Most freq.		*10.06*	*25.00*	*14.35*

Table 5
Results on the test set for the *special track*, subtask B, divided by area. Baselines are italicized and highlighted in *yellow*.

	Team	Run	Avg dist (km)
	Gallo-Italic area		
	Centroid		*102.01*
	kNN		*102.16*
1	SCG	(1)	102.41
2	SCG	(2)	104.59
3	SCG	(3)	111.79

Subtask B: Fine-grained geolocation In Table 5, we report the results for the area chosen by participants in subtask B. As for subtask A, we however do not have enough information to discuss further the *SCG*'s results.

6. Analysis and discussion

In this section, we analyze the approaches adopted by teams along several dimensions, providing a discussion and the insights derived from the shared task.

Models Apart from *SCG*, all participant teams used transformer-based language models for their runs. *Salogni* adopted an Italian RoBERTa-based model. *DANTE* and *baρtti* used versions of BERT pre-trained on Italian data, with the former using a much larger pre-training corpus than the latter, which might have impacted on the *DANTE* runs ranking first in subtask A. In contrast, *galliz* employed an English pre-trained BERT model, which still outperformed the logistic regression baseline in subtask A for both the tracks. This might indicate that subword tokenization in these models is suboptimal for the language varieties in DiatopIt, which naturally exhibits many non-Italian tokens with varied written forms, resulting in potentially small differences between Italian and English pre-trained models. Lastly, *extremITA* used a T5-based model pre-trained on Italian data and a LLaMA-based instruction-tuned model. Their results showed that recent large language models fine-tuned on disparate tasks are still far from tackling tasks such as GeoLingIt.

Multi-task learning Both *baρtti* and *DANTE* used multi-task learning in their submissions. While *baρtti* employed it during fine-tuning to exploit subtask A information to tackle subtask B and vice versa, *DANTE* used multi-task learning during a further stage of pre-training of a BERT-based model pre-trained on Italian data, which was then used to separately fine-tuning it on subtask A and B. Their pre-training setup consists of four tasks, including region-informed objectives, such as the prediction of the provenance region of posts and tokens. The approach followed by *DANTE* appears to lead to better performance in subtask A, whereas jointly training on both subtasks as done by *baρtti* seems to help in modeling fine-grained geolocation. Future work may shed light on how those approaches can help each other.

External resources Some participants used external resources to integrate the available data for the task. Three teams (i.e., *DANTE*, *galliz*, and *baρtti*) used data from a website containing a series of stories, poems, idioms, recipes, and articles in different language varieties that are spoken across Italy (i.e., Dialettando). In addition to this, *DANTE* also leveraged Wikipedia articles written in some of the language varieties that are present in our data. Both *DANTE* and *baρtti* used additional data from the Italian Wikipedia. For the *special track*, *galliz* also used lemmas from both a vocabulary of contemporary Florentine and a webpage for the Romanesco dialect (cf. Section 5.2.2). While *galliz* and *baρtti* used external data to create vocabularies, *DANTE* used it for pre-training their models. All of the teams who used external resources outperformed both baselines in both tasks, signaling that the use of external resources may indeed be pivotal in tackling the GeoLingIt task.

Data augmentation *baρtti* and *galliz* employed data augmentation techniques in order to artificially increase the amount of training data. *galliz* used external data to fine-tune an Italian word embeddings model, and then exploited it to swap randomly selected tokens with other semantically close ones. The *baρtti* team, on the other hand, constructed a vocabulary using external resources and then used it to randomly substitute tokens with other tokens from the vocabulary. Both teams outperformed our baselines, showing that the augmentation and diversification of training data can be useful for the task.

7. Conclusions

This paper provided an overview of GeoLingIt, the first shared task focused on geolocation of linguistic variation in Italy. The task attracted wide interest from the community, registering 37 expressions of interest and 35 official runs. After presenting participants' results and the adopted approaches, we outlined the main insights from the shared task. Besides natural language processing, we hope that GeoLingIt sensitized the community on the linguistic diversity of the country.

References

[1] M. Maiden, M. Parry, The dialects of Italy, Romance Linguistics, Routledge, 1997.

[2] A. Ramponi, NLP for language varieties of Italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022). URL: https://arxiv.org/abs/2209.09757.

[3] F. Avolio, Lingue e dialetti d'Italia, Le Bussole, Carocci, Roma, Italy, 2009.

[4] J. Eisenstein, What to do about bad language on the internet, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 359–369.

[5] R. van der Goot, A. Ramponi, A. Zubiaga, B. Plank, B. Muller, I. San Vicente Roncal, N. Ljubešić, Ö. Çetinoğlu, R. Mahendra, T. Çolakoğlu, T. Baldwin, T. Caselli, W. Sidorenko, MultiLexNorm: A shared task on multilingual lexical normalization, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, Online, 2021, pp. 493–509. URL: https://aclanthology.org/2021.wnut-1.55. doi:10.18653/v1/2021.wnut-1.55.

[6] G. Berruto, Quale dialetto per l'Italia del duemila? Aspetti dell'italianizzazione e risorgenze dialettali in Piemonte (e altrove), in: Lingua e dialetto nell'Italia del Duemila, Congedo, 2006, pp. 101–127.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.

[9] B. Han, A. Rahimi, L. Derczynski, T. Baldwin, Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text, in: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 213–217.

[10] M. Gaman, D. Hovy, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, C. Purschke, Y. Scherrer, M. Zampieri, A report on the VarDial evaluation campaign 2020, in: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 1–14.

[11] B. R. Chakravarthi, G. Mihaela, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, M. Zampieri, Findings of the VarDial evaluation campaign 2021, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Kiyv, Ukraine, 2021, pp. 1–11.

[12] G. B. Pellegrini, Carta dei dialetti d'Italia, Profilo dei Dialetti Italiani, Pacini, Pisa, Italy, 1977.

[13] A. Ramponi, S. Tonelli, Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3027–3040. URL: https://aclanthology.org/2022.naacl-main.221. doi:10.18653/v1/2022.naacl-main.221.

[14] A. Koudounas, F. Giobergia, I. Benedetto, S. Monaco, L. Cagliero, D. Apiletti, E. Baralis, baρtti at GeoLingIt: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] G. Gallipoli, M. La Quatra, D. Rege Cambrin, S. Greco, L. Cagliero, DANTE at GeoLingIt: Dialect-aware multi-granularity pre-training for locating tweets within Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[16] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[17] T. Labruna, S. Gallo, Galliz at GeoLingIt: Enhancing BERT with vocabulary knowledge for predicting the region of language varieties of Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[18] I. Salogni, Salogni at GeoLingIt: Geolocalization by fine-tuning BERT, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.