Towards a Foundation Model for Geospatial Artificial Intelligence (Vision Paper)

Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, Stefano Ermon
[1]Department of Computer Science, Stanford University, Stanford, CA, USA
{maigch,cundy,kechoi,ermon}@cs.stanford.edu
[2]Department of Geography, University at Buffalo, Buffalo, NY, USA
yhu42@buffalo.edu
[3]Google, Mountain View, CA, USA
nlao@google.com

ABSTRACT

Large pre-trained models, also known as *foundation models* (FMs), are trained in a task-agnostic manner on large-scale data and can be adapted to a wide range of downstream tasks by fine tuning, few-shot, or even zero-shot learning. Despite their successes in language and vision tasks, we have yet to see an attempt to develop foundation models for geospatial artificial intelligence (GeoAI). In this work, we explore the promises and challenges for developing multimodal foundation models for GeoAI. We first show the advantages of this idea by testing the performance of existing Large pre-trained Language Models (LLMs) (e.g. GPT-2 and GPT-3) on two geospatial semantics tasks. Results indicate that these task-agnostic LLMs can outperform task-specific fully-supervised models on both tasks with 2-9% improvement in a few-shot learning setting. However, we also show the limitations of these existing foundation models given the multimodality nature of GeoAI, especially when dealing with geometries in conjunction with other modalities. So we discuss the possibility of a multimodal foundation model which can reason over various types of geospatial data through geospatial alignments. We conclude this paper by discussing the unique risks and challenges to develop such model for GeoAI.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Unsupervised learning*; • **Applied computing** → **Earth and atmospheric sciences**.

KEYWORDS

Foundation models, large language models, geospatial artificial intelligence

ACM Reference Format:
Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, Stefano Ermon. 2022. Towards a Foundation Model for Geospatial Artificial Intelligence (Vision Paper). In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557915.3561043

SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9529-8/22/11.
https://doi.org/10.1145/3557915.3561043

```
[Instruction] ...
Paragraph: Alabama State Troopers say a Greenville man has died of his injuries
    ↪ after being hit by a pickup truck on Interstate 65 in Lowndes County.
Q: Which words in this paragraph represent named places?
A: Alabama; Greenville; Lowndes
...
--
Paragraph: The Town of Washington is to what Williamsburg is to Virginia.
Q: Which words in this paragraph represent named places?
A: Washington; Williamsburg; Virginia
```

Listing 1: Typonym recognition with LLMs, e.g., GPT-3. Yellow block: the text snippet to be annotated. Orange box: GPT-3 outputs. 8 few-shot samples are used in this prompt. We only show 1 here while skipping others with "..." to save space.

```
[Instruction] ...
Paragraph: Papa stranded in home. Water rising above waist. HELP 8111 Woodlyn Rd
    ↪ , 77028 #houstonflood
Q: Which words in this paragraph represent location descriptions?
A: 8111 Woodlyn Rd, 77028
...
--
Paragraph: HurricaneHarvey Help Need AT 7506 Jackrabbit Rd, Houston, TX 77095.
Q: Which words in this paragraph represent location descriptions?
A: 7506 Jackrabbit Rd, Houston, TX 77095
```

Listing 2: Location description recognition with LLMs, e.g., GPT-3. Yellow block: the input text snippet. Orange box: GPT-3 outputs. 11 few-shot samples are used while 1 is shown.

1 INTRODUCTION

Recent trends in machine learning (ML) and artificial intelligence (AI) speak to the unbridled powers of data and compute – extremely large models trained on Internet-scale datasets such as GPT-3[5], CLIP [22], and DALL·E2 [23] have achieved state-of-the-art (SOTA) performance on a diverse range of learning tasks. In particular, their unprecedented success has spurred a *paradigm shift* in the way that modern-day ML models are trained. Rather than learning task-specific models from scratch [8, 13, 26], such pre-trained models (so-called "foundation models (FMs)" [4]) are *adapted* via fine-tuning or few-shot/zero-shot learning strategies and subsequently deployed on a wide range of domains [5, 22]. Such FMs allow for the transfer and sharing of knowledge across domains, and mitigate the need for task-specific training data.

Despite their successes, there exists very little work exploring the development of an analogous *FM for geospatial artificial intelligence (GeoAI)*. The key technical challenge here is the inherently *multimodal* nature of GeoAI. The core data modalities in GeoAI include text, images (remote sensing or streetview images), knowledge

graphs, and geospatial vector data, all of which contain important geospatial information (e.g. geometries). Each modality exhibits special structures that require its own unique representation – effectively combining all of these representations together with the appropriate inductive biases within a single model requires careful design. This limitation prevents a straightforward application of existing pre-trained FMs across all GeoAI tasks.

In this paper, we lay the groundwork for developing FMs for GeoAI. First, to showcase their potential for GeoAI, we demonstrate the advantages of LLMs over existing baselines on several well-defined geospatial semantics tasks (See Listing 1 and 2). Next, we detail the challenges for developing FMs for GeoAI. Since creating one FM for all GeoAI data modalities can be very difficult, we start this discussion by examining the possibility of developing FMs for GeoAI tasks that share one data modality. Then, we propose our vision for a novel, multimodal FM framework for GeoAI that addresses the previous challenges. Finally, we point out some potential risks and challenges that must be considered when developing such general-purpose models for GeoAI.

2 PRE-TRAINED LANGUAGE MODELS HOLD PROMISE FOR GEOAI

As a starting point for our discussion, we demonstrate empirically the promise of leveraging LLMs for solving geospatial semantics tasks. Our result not only demonstrates the effectiveness of such general-purpose, few-shot learners in the geospatial semantics domain, but also challenges the current paradigm of training task-specific models as a common practice in GeoAI research.

We compare the performance of 4 pre-trained GPT-2 [21] models of varying sizes as well as the most recent GPT-3 [5] model with multiple *supervised, task-specific* baselines on two representative geospatial semantics tasks: (1) toponym recognition [8, 25], and (2) location description recognition [9]. As a subtask of named entity recognition (NER), the goal of toponym recognition is to recognize named places from a given text snippet. The location description recognition task is slightly more challenging – given a text snippet (e.g., tweets), the goal is to recognize more fine-grained location descriptions such as home addresses, highways, roads, and administration regions. We use the Hu2014 [10] and Ju2016 [11] benchmark datasets for the first task and HaveyTweet2017 [9] for the second task. We adapt 5 pre-trained GPT models to perform both tasks by using appropriate prompts containing few-shot training examples. Listing 1 and 2 shows the prompts used for both tasks.

Table 1 compares GPT-2/GPT-3 with 13 baselines on these three datasets. With the exception of the smallest GPT-2 model, all other LLMs consistently outperform the fully-supervised baselines on the Hu2014 dataset, even though they only require a small set of natural language instructions without any additional training. GPT-3 in particular demonstrated an 8.7% performance improvement over the previous SOTA (TopoCluster [6]). On the Ju2016 dataset, we found that GPT-2-XL outperforms the previous SOTA (NeuroTPR [26]) by over 2.5% while using only *8 few-shot examples in the prompt*. In contrast, a task-specific model, e.g., NeuroTPR, needs to be supervised trained on 599 labelled tweets and labelled sentences generated from 3000 Wikipedia articles. In accordance with existing empirical findings [5, 21], we also found that the performance of these LLMs tended to scale with the the number of learnable parameters. On

Table 1: Geospatial semantics result for various GPT models and baselines. (A) General NER; (B) No Neural Network (NN) based geoparsers; (C) Fully-supervised NN-based geoparsers; (D) Few-show learning with LLMs. "(#Param)" indicates the number of learnable parameters of LLMs. The results of all baselines are obtained from [25] and [26] except "0.675†", which is obtained by rerunning the official code of [26]. *We evaluate GPT-3 on randomly sampled 544 Ju2016 examples (10% of the dataset), because of the GPT-3 API limitation.

	Model (#Params)	Typonym Recog.		Location Desc. Recog.		
		Hu2014[10]	Ju2016[11]	HaveyTweet2017[9]		
		Accuracy	Accuracy	Precision	Recall	F-Score
A	Stanford NER (nar. loc.) [26]	0.787	0.010	**0.828**	0.399	0.539
	Stanford NER (bro. loc.) [26]	-	0.012	0.729	0.440	0.548
	Retrained Stanford NER [26]	-	0.078	0.604	0.410	0.489
	spaCy NER (nar. loc.) [26]	0.681	0.000	0.575	0.024	0.046
	spaCy NER (bro. loc.) [26]	-	0.006	0.461	0.304	0.366
	DBpedia Spotlight [20]	0.688	0.447	-	-	-
B	Edinburgh [2]	0.656	0.000	-	-	-
	CLAVIN [25]	0.650	0.000	-	-	-
	TopoCluster [6]	0.794	0.158	-	-	-
C	CamCoder [8]	0.637	0.004	-	-	-
	Basic BiLSTM+CRF [14]	-	0.595	0.703	0.600	0.649
	DM_NLP (top. rec.) [27]	-	0.723	0.729	0.680	0.703
	NeuroTPR [26]	0.675†	0.821	0.787	0.678	**0.728**
D	GPT-2 [21] (117M)	0.556	0.650	0.540	0.413	0.468
	GPT-2-Medium [21] (345M)	0.806	0.802	0.529	0.503	0.515
	GPT-2-Large [21] (774M)	0.813	0.779	0.598	0.458	0.518
	GPT-2-XL [21] (1558M)	0.869	**0.846**	0.492	0.470	0.481
	GPT-3 [5] (175B)	**0.881**	0.811*	0.603	**0.724**	0.658

the HaveyTweet2017 dataset, GPT-3 achieves the best recall score across all methods. However, all LLMs have rather low precision (and therefore low F1-scores). This is because LLMs implicitly convert the location description recognition problem into a natural language generation problem (see List 2), meaning that they are not guaranteed to generate tokens that appear in the input text. This result clearly showcases the potential of LLMs to dramatically reduce the need for customized architectures or large labelled datasets.

3 A MULTIMODAL FM FOR GEOAI

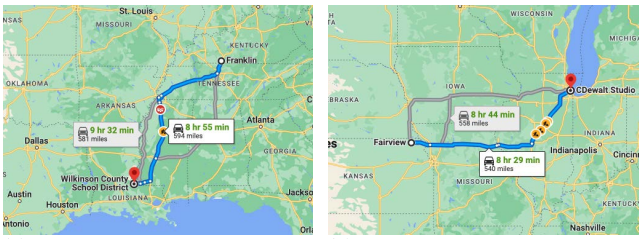
Although LLMs exhibit strong performance on several geospatial semantics tasks, they are unable to handle the wide range of data modalities presented in GeoAI. In this section, we discuss the challenges unique to each data modality, then propose a potential framework for future GeoAI which leverages a multimodal FM.

3.1 Text and Geospatial Semantics

Despite the promising results showed in Table 1, LLMs still struggle with more complex geospatial semantics tasks such as toponym resolution/geoparsing [2, 8, 25] and geographic question answering (GeoQA) [17, 18], since LLMs are unable to perform (implicit) spatial reasoning in a way that is grounded in the real world.

As a concrete example, we illustrate the shortcomings of GPT-3 on a geoparsing task. Using two examples from the Ju2016 dataset, we ask GPT-3 to both: 1) recognize toponyms; and 2) predict their geo-coordinates. The prompt is shown in List 3 while the geoparsing results are visualized in Figure 1. We see that in both cases, GPT-3 can correctly recognize the toponyms but the predicted coordinates are 500+ miles away from the ground truth. Moreover, we notice that with a small spatial displacement of the generated geo-coordinates, GPT-3's log probability for this new coordinates decreases significantly. In other words, the probability of coordinates

generated by the LLM does not follow the First Law of Geography. GPT-3 also generates invalid latitudinal/longitudinal coordinates, indicating that existing LLMs are still far from gracefully handling complex numerical and spatial reasoning tasks.



(a) [TEXT]: Franklin is a city in and the county seat of simpson county, ...

(b) [TEXT]: the city of Fairview had a population of 260 as of july 1, 2015. ...

Figure 1: Geoparsing examples of GPT-3 on the Ju2016 dataset comparing the predicted coordinates (dropped pins) and the ground truth coordinates (starting points). The recognized toponyms are underlined in text.

```
[Instruction] ...
Paragraph: San Jose was founded in 1803 when allotments of land were made ...
Q: Which words in this paragraph represent named places?
A: San Jose; New Mexico

Q: What is the location of San Jose?
A: 35.39728, -105.47501
...
--
Paragraph: the city of fairview had a population of 260 as of july 1, 2015. ...
Q: Which words in this paragraph represent named places?
A: Fairview

Q: What is the location of Fairview?
A: 41.85003, -87.65005
```

Listing 3: Geoparsing with LLMs, e.g., GPT-3. Yellow block: the text snippet to be geoparsed. Orange box: GPT-3 outputs.

3.2 Remote Sensing

With the advancement of computer vision technology, deep vision models have been successfully applied to different kinds of remote sensing (RS) tasks including image classification/regression [3, 24], land cover classification [3], and object detection[13]. Unlike the usual vision tasks which usually work on RGB images, RS tasks are based on multispectral/hyperspectral images from different sensors. Most existing RS works focus on training one model for a specific RS task and a specific sensor [13]. However, we see the trend of FMs in the CV field such as CLIP [22] to be further developed to meet the unique challenges of remote sensing tasks.

Aside from being **task-agnostic**, the desiderata for a remote sensing FM include being: 1) **sensor-agnostic**: it can seamlessly reason among RS images from different sensors with different spatial or spectral resolutions; and 2) **spatiotemporally-aware**: it can handle the spatiotemporal metadata of RS images and perform geospatial reasoning for tasks such as image geolocalization and object tracking. Recent developments here include geography-aware RS models [3] or self-supervised/unsupervised RS models [3, 24], all of which are task-agnostic. However, we have yet to develop a FM for RS tasks which can satisfy all such properties.

3.3 Geospatial Vector Data

Another critical challenge in developing FMs for GeoAI is the complexity of geospatial vector data. In contrast with NLP and CV where text (1-D) or images (2-D) are well-structured and more suitable to common neural network architectures, GeoAI vector data exhibits more complex data structures in the form of points, polylines, polygons, and networks [19]. So it is particularly challenging to develop a FM which can seamlessly encode or decode different kinds of vector data.

Noticeably, each vector data format is constructed based on locations. So recently developed location encoding techniques [16, 19] can be seen as a fundamental building block for such a model. Moreover, since encoding (e.g., geo-aware image classification[16]) or decoding (e.g., geoparsing [25]) geospatial vector data, or conducting spatial reasoning (e.g., GeoQA [18]) is an indispensable component for most GeoAI tasks, developing FMs for vector data is the key step towards a multimodal FM for GeoAI. This point also differentiates GeoAI FMs from existing FMs in other domains.

3.4 A Multimodal FM for GeoAI

Except for those three data types, there are also other datasets frequently studied in GeoAI such as streetview images, geo-tagged videos, and geospatial knowledge graphs. Given all these diverse data modalities, the question is how to develop a multimodal FM for GeoAI that best integrates all of them.

When we take a look at the existing multimodal FMs such as CLIP [22], DALL·E2 [23], MDETR [12] and VATT [1] we can see the following general architecture: 1) **starting with separate embedding modules to encode different modalities of data** (e.g., a Transformer for texts and ResNet50 for images [22]); 2) (optionally) **mixing the representations** of different modalities by concatenation; 3) (optionally) **more Transformer layers** for across modality reasoning, which can achieve certain degree of alignment based on semantics, e.g., the word "hospital" attending to a picture of hospital; 4) **generative or discriminative prediction modules** for different modalities to achieve self-supervised training.

One weak point of these architectures is the lack of integration with vector data, which is the backbone of spatial reasoning and helps alignment among multi-modalities in GeoAI. This is considered central and critical for GeoAI tasks. Therefore, we propose to replace step 2 with **aligning the representations** of different modalities (e.g., geo-tagged texts and RS images) by augmenting their represenations with location encoding[16] before mixing them. Figure 2 illustrates this idea. Geo-tagged text data and remote sensing (or streetview) images can be easily aligned via their geographic footprints (vector data). The key advantages of such model are to enable spatial reasoning and knowledge transfer across modalities.

4 RISKS AND CHALLENGES

Geographic Bias. It is well known that foundation models have the potential to amplify existing societal inequalities and biases present in the data [4]. A key consideration for GeoAI in particular is *geographic bias* [15], which is often overlooked by AI research. For example, Zilong et al. [15] showed that all current geoparsers are highly geographically biased towards data-rich regions. Compared to task-specific models, FMs suffer more from geographic bias since: 1) the training data is collected in large-scale which is likely to be

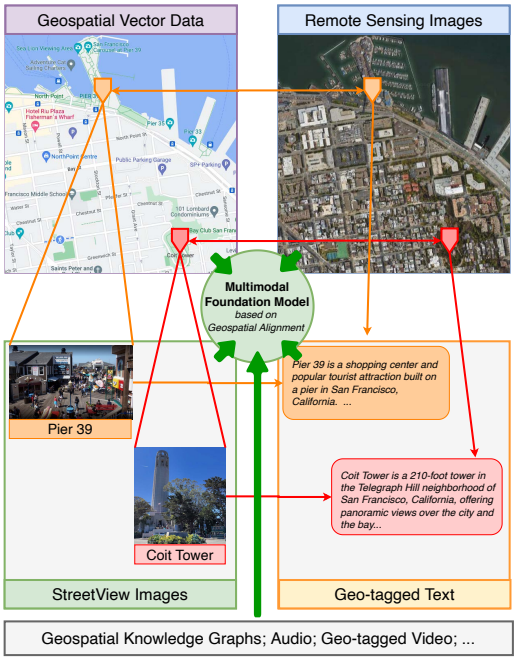


Figure 2: A multimodal FM which achieves alignment among different data sources via their geospatial relationships.

dominated by overrepresented communities or regions; 2) the huge number of learnable parameters and complex model structures make model interpretation and debiasing much more difficult; 3) the geographic bias of the FMs can be easily inherited by all the adapted models downstream [4], and thus bring much more harm to the society. This indicates a pressing need for designing proper (geographic) debiasing frameworks.

Spatial Scale. Geographic information can be represented in different spatial scales, which means that the same geographic phenomenon/object can have completely different spatial representations (points vs. polygons) across GeoAI tasks. For example, an urban traffic forecasting model must represent San Francisco (SF) as a complex polygon, while a geoparser usually represents SF as a single point. Since FMs are developed for a diverse set of downstream tasks, they need to be able to handle geospatial information with different spatial scales, and infer the right spatial scale to use given a downstream task. Developing such a module is a critical component for an effective GeoAI FM.

Generalizability v.s. Spatial Heterogeneity. An open problem for GeoAI is how to achieve model generalizability ("replicability" [7]) across space while still allowing the model to capture spatial heterogeneity. Given geospatial data with different spatial scales, we desire a FM that can learn general spatial trends while still memorizing location-specific details. Will this generalizability introduce unavoidable intrinsic model bias in downstream GeoAI tasks? Will this memorized localized information lead to an overly complicated prediction surface for a global prediction problem? With large-scale training data, this problem can be amplified and requires care .

5 CONCLUSION

In this paper, we discuss the promises and challenges for developing multimodal foundation models (FMs) for GeoAI. The superiority of FMs is demonstrated by comparing the performance of existing LLMs as few-shot learners with fully-supervised task-specific SOTA models on two geospatial semantics tasks. We then propose our vision for a novel multimodal FM for GeoAI. We conclude this work by discussing some unique challenges and risks for such model.

Acknowledgement: This work is supported by the NSF (Grant No. BCS-2117771). SE, CC, and KC acknowledge support by NSF (#1651565, #1522054, #1733686), ONR (N00014-19-1-2145), AFOSR (FA9550-19-1-0024), ARO (W911NF2110125), CZ Biohub, Amazon AWS, and Sloan Fellowship. KC is supported by the Qualcomm Innovation Fellowship and Two Sigma PhD Diversity Fellowship.

REFERENCES

[1] Hassan Akbari et al. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *NeurIPS 2021*, Vol. 34. 24206–24221.
[2] Beatrice Alex et al. 2015. Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing* 9, 1 (2015).
[3] Kumar Ayush et al. 2021. Geography-aware self-supervised learning. In *CVPR 2021*. 10181–10190.
[4] Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
[5] Tom Brown et al. 2020. Language models are few-shot learners. *NIPS 2020* 33 (2020), 1877–1901.
[6] Grant DeLozier, Benjamin Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a novel geolocation corpus from historical texts. In *LAW-X 2016*. 188–198.
[7] Michael F Goodchild and Wenwen Li. 2021. Replication across space and time must be weak in the social and environmental sciences. *PNAS* 118, 35 (2021).
[8] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which Melbourne? Augmenting Geocoding with Maps. In *ACL 2018*. 1285–1296.
[9] Yingjie Hu, , et al. 2020. How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey. In *GIScience 2020*.
[10] Yingjie Hu et al. 2014. Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In *GIR Workshop 2014*. 1–8.
[11] Yiting Ju et al. 2016. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *EKAW 2016*. Springer, 353–367.
[12] Aishwarya Kamath et al. 2021. MDETR–Modulated Detection for End-to-End Multi-Modal Understanding. *arXiv preprint arXiv:2104.12763* (2021).
[13] Darius Lam et al. 2018. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856* (2018).
[14] Guillaume Lample et al. 2016. Neural Architectures for Named Entity Recognition. In *NAACL-HIT 2016*. 260–270.
[15] Zilong Liu et al. 2022. Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing. *AGILE 2022* 3 (2022), 1–13.
[16] Gengchen Mai et al. 2020. Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells. In *ICLR 2020*. openreview.
[17] Gengchen Mai et al. 2020. SE-KGE: A Location-Aware Knowledge Graph Embedding Model for Geographic Question Answering and Spatial Semantic Lifting. *Transactions in GIS* (2020).
[18] Gengchen Mai et al. 2021. Geographic question answering: challenges, uniqueness, classification, and future directions. *AGILE 2021* 2 (2021), 1–21.
[19] Gengchen Mai et al. 2022. A review of location encoding for GeoAI: methods and applications. *IJGIS* 36, 4 (2022), 639–673.
[20] Pablo N Mendes et al. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-Semantics 2011*.
[21] Alec Radford et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[22] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *ICML 2021*. PMLR, 8748–8763.
[23] Aditya Ramesh et al. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
[24] Esther Rolf et al. 2021. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications* 12, 1 (2021), 1–11.
[25] Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS* 23, 6 (2019), 1393–1419.
[26] Jimin Wang, Yingjie Hu, and Kenneth Joseph. 2020. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS* 24, 3 (2020), 719–735.
[27] Xiaobin Wang et al. 2019. DM_NLP at semeval-2018 task 12: A pipeline system for toponym resolution. In *SEMEVAL 2019*. 917–923.